

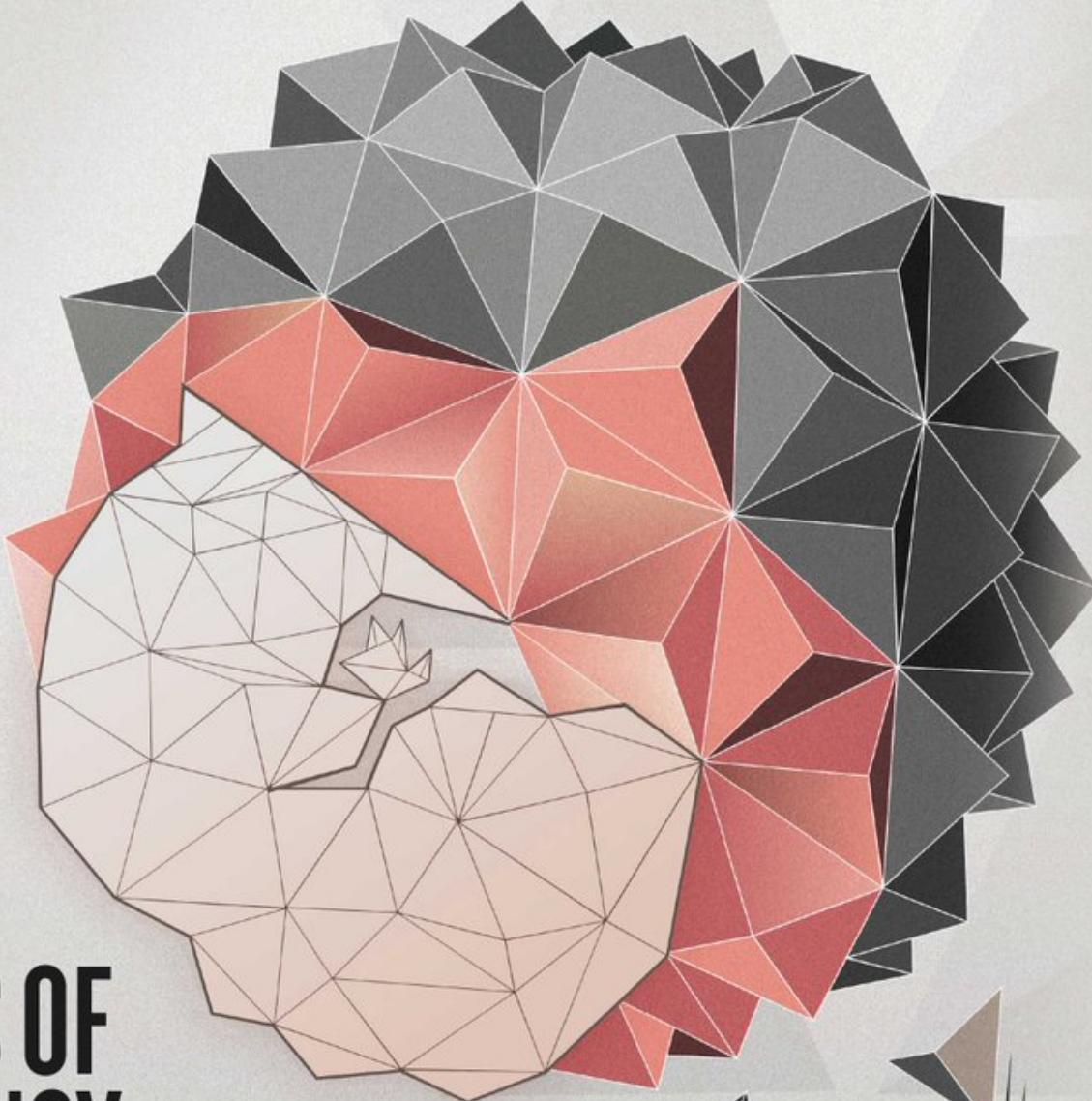
nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

*The unique
maternal cell
types that
support the
developing
placenta*

PAGES 337 & 347

SECRETS OF PREGNANCY



MEDICAL RESEARCH

THE POWER OF FAILURE

Access to negative results
can speed drug development

PAGE 317

ASTRONOMY

SIGNALS FROM THE SNOW LINE

Candidate cold super-Earth
found orbiting Barnard's star

PAGES 329 & 365

CHEMISTRY

PRACTICAL SYNTHESIS

A simple route to forging
carbon–carbon bonds

PAGES 336 & 379

NATURE.COM

15 November 2018

Vol. 563, No. 7731

THIS WEEK

EDITORIALS

POLITICS Science could benefit from US election results **p.294**

WORLD VIEW Point out the roads that will take us nowhere **p.295**



IN THE BALANCE Thinning snow threatens sports and water supplies **p.297**

A boost for Palestinian science

Researchers from around the world can help to support and collaborate with colleagues in troubled regions.

Scientist-statesman Chaim Weizmann was Israel's first president. He was also a chemist, and had earlier helped to create the world-ranking research institute in Rehovot that now bears his name. He famously declared that a state should be built on science — then he tried to do just that. Today, Israel is a global player, investing more than 4% of its gross domestic product in research and development, one of the highest proportions in the world. Its scientists are winning some of the most prestigious international grants, and are collaborating with partners around the world.

Yet since 2002, scientists in Israel have faced calls for an academic boycott, a move that *Nature* opposed at the time and continues to disagree with (see *Nature* 417, 1; 2002). The controversy is a reminder that Israel's stellar science is often overshadowed by the troubled politics of the region. Not least of these is Israel's occupation of the Palestinian territories, which concerns researchers around the world and many Israeli scientists, too.

If Israeli science is overshadowed by politics, then Palestinian science has an even more diminished profile. A *Nature* reporting trip to the region showed how hard conditions are. Frustration and anger bubble below the surface — but the scientific spirit endures.

In the West Bank and East Jerusalem, the welcome seeds of sustainable research are being planted. Wittingly or not, academics there poignantly echo Weizmann's sentiment that any new state needs to be based on science. The challenges they face are immense (see page 308). The conditions of the occupation prevent people from moving freely in and out of the territories, and block or delay imports of crucial research consumables. Money for research, from any source, is at best a pittance.

The situation in Gaza is even more critical. There, even the basics of daily life, such as continuous access to drinking water, cannot be taken for granted. And scientists have few resources for research and even less contact with the outside world. Electricity is rationed to a few hours per day. Yet science is pursued. Physicist Hala El-Khözondar at the Islamic University of Gaza, barely 70 kilometres from Tel Aviv, has started to use a recycled truck battery to charge her laptop so she can achieve a full working day.

The Weizmann Institute of Science did not have it easy, either, in its early days, when the newly founded and very poor state of Israel had to build itself up from scratch and defend itself from hostile neighbours. International support was fundamental in the country's rise to a scientific power.

International support can help Palestinians to at least establish a scientific base. There is already a scattering of foreign funding programmes and external collaborations, and more would help. Scientists can help by donating equipment. And some universities generously grant online access to their own libraries for short periods. When that happens, says El-Khözondar, she and her colleagues binge read. "The environment affects everything, even your mood and certainly your research," she says. "We want to be up to date for when the situation

changes; so we do our best to keep up."

Interaction with scientists abroad can reduce the isolation felt by Palestinian colleagues, as well as open up more opportunities for funding and collaboration. Many Palestinian scientists don't want to work with Israelis, saying that it would normalize the occupation. Few, in any case, dare. They say the mood in the territories is so bitter that if they openly collaborate with an Israeli team, they risk their lab — or worse, their home — being torched. This is tragic.

"International support can help Palestinians to at least establish a scientific base."

Science is not going to solve the Israeli-Palestinian dispute, but it can be helpful in keeping international dialogue open. Time and again, scientific diplomacy has proved a useful tool in broad efforts to resolve disputes. The classic example is how CERN, the international particle-physics lab near Geneva, Switzerland, signed agreements with Soviet scientific institutes at the height of the cold war in the 1960s. The Synchrotron-Light for Experimental Science and Applications in the Middle East (SESAME) facility in Jordan, which was inaugurated last year, was designed for a similar purpose and includes both Israeli and Palestinian researchers. Through such steps, science can help in building a state of mind, and maybe more. ■

Mutual benefit

Researchers should do much more to involve those who take part in clinical trials.

When researchers at the drug giant Pfizer wanted to improve their clinical trials, the people who had taken part had a clear suggestion: researchers should say thank you.

It is a simple request, but a revealing one. When a clinical trial is completed, many participants walk away empty-handed. Most never hear from the investigators or the trial's sponsor again. Many do not learn the results of the study in which they took part. It's not good enough — and it indicates a deeper problem.

As we discuss in a News Feature on page 312, clinical-trial participants and the people who care for them are increasingly seen as partners in research. They are more informed than ever about their conditions and their medical options. And they are demanding — and receiving — more of a say in how clinical trials are designed and conducted. Some of this activity has been boosted by social media, which has allowed people with medical conditions and their carers to band together, share their experiences and advocate for change.

There has been some progress. In the late 1990s, Sharon Terry

arrived for a meeting at the US National Institutes of Health to discuss a project to study a condition called pseudoxanthoma elasticum, which affects elastic fibres in some tissues. Even though Terry was the founder of an advocacy group that wanted to fund part of the study, she was told that she could not join the meeting because she wasn't trained in biomedical research. She was eventually allowed to attend, but only if she served as an assistant to the medical director of her group. (Terry decided to pull funding for the project.)

Terry says it is hard to imagine the same scene today. Many pharmaceutical companies and medical centres now routinely consult people with a condition about clinical-trial designs, to get early feedback rather than risk launching a trial that no one wants to join. In response, trial organizers have tweaked protocols and created research programmes. In cancer studies, for example, this type of feedback has fuelled a push to find ways to combat the side effects of cancer treatment, and to improve care for survivors of cancer.

The benefits of such an approach are persuasive. Closer engagement with participants could yield clinical-trial protocols that are more effective and convenient for patients. This can translate into a trial that meets its enrolment targets more quickly, and which has a lower dropout rate.

Nancy Roach, founder of the advocacy group Fight Colorectal Cancer in Springfield, Missouri, recalls a meeting at the US National Cancer Institute about a trial in which participants would be assigned a treatment on the basis of their tumour mutations. An early proposal called for tumour samples to be characterized in three to four weeks. Roach, as well as others at the meeting who represented the participants, said it would not work: the longest they would be willing to wait before settling

on a course of treatment was two weeks. After a subsequent survey of clinicians and investigators confirmed that they would also wait only two weeks before deciding on a treatment, the project team worked with pathologists at the trial sites to shorten the time it took to process the samples. The trial, called NCI-MATCH, initially had trouble meeting those goals because so many more people enrolled in the study than expected. So far, there are more than 6,000 participants.

"It is important to make sure that patient engagement is backed by meaningful action."

More projects should follow this approach. As the phrase 'patient engagement' sweeps through medical science, it is important to make sure that it's backed by meaningful action. It is not enough to put a potential trial participant in the room during meetings to discuss protocol designs. And it's unacceptable that some scientists still consult people about a trial protocol only after it has been

approved by a review board, when all involved are reluctant to revise it.

Engagement means offering training to participants and their carers so that they have the skills to contribute with confidence. Some say that it is intimidating to be in a room full of specialists, with the added responsibility of speaking for an entire community of people who have a medical condition. Engagement is also about researchers being willing to incorporate patient feedback. There are plenty of examples of best practice to follow, including lessons from social scientists who have studied community engagement to learn how best to achieve it.

Clinical trials depend on the willingness of participants, some of whom are critically ill. They all deserve a thank you. They rightly expect much more. ■

Welcome change

Science-based policies should benefit from midterm election results in the United States.

US President Donald Trump has taken a wrecking ball to the climate and environment policies of his predecessor, Barack Obama, over the past two years. To some extent, this is to be expected: any administration has the ability and right to lay out its policies and set a new course. But the Trump administration has also shown a complete disregard for the science and evidence that should underpin policy decisions.

In many cases, Republicans in Congress have been all too happy to sit back and watch. The political dynamic will now change, given that Democrats took control of the House of Representatives in the midterm elections last week (see page 302).

As *Nature* went to press, officials were still tallying votes in several close races, but the new balance of power is clear. Democrats have so far picked up 32 seats in the House, giving them a slim but significant majority they can use to block the administration's legislative agenda — just as Republicans did when Obama was president. The Trump administration has often used its executive authorities to advance its agenda independently of Congress, and will surely continue to do so. The difference now is that Democrats will have the power to investigate and raise questions about policies, and to issue subpoenas to compel testimony from reluctant administration officials. This won't necessarily stop the administration, but it will put a public spotlight on the decision-making process. For anybody who cares about evidence-based policies — including this journal — this is good news.

It's a different situation in the Senate, where Republicans will pick up at least two seats. Given the current polarization between Democrats and Republicans, the odds of bipartisanship cooperation are slim, but there are some areas in which the two parties might work together. One is the protection of funding for science and science-based

agencies: the current Republican-led Congress has already declined Trump's demands to slash funding for the Environmental Protection Agency (EPA) and other such groups, and there will be little appetite to do so next year. (The long-term budget outlook is bleak, so there might still be plenty of cuts to come.) The other point on which the two parties could unite is spending for research infrastructure.

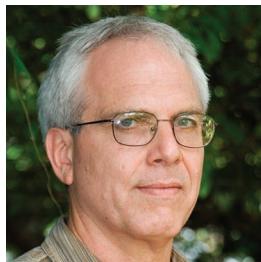
When it comes to science, all eyes are now on changes to the committees that oversee health and environmental agencies — most notably the EPA, a primary target of Trump's scorn and the main vehicle for his efforts to dismantle rules and regulations that protect the environment and public health but burden industry.

At minimum, expect a change in the language around global warming. The current chair of the House Committee on Science, Space, and Technology, which regularly weighs in on scientific and technical issues, has repeatedly questioned climate science while launching investigations into alleged wrongdoing by scientists and scientific agencies. But Democrat Eddie Bernice Johnson, who is a registered nurse and now the probable future chair of the committee, plans to set the record on climate change straight in hearings next year, starting with an acknowledgement that "it is real".

As Democrats push back, legal battles will continue to play out in the courts. Republican gains in the Senate will make it even easier for the administration to appoint judges and push the judicial system in a conservative direction. But federal judges have already rejected some of Trump's decisions for lack of scientific analysis. Last week, a federal district court blocked construction of the Keystone XL pipeline, which would help to transport crude oil from the Canadian tar sands to the United States; the court ruled that the administration had "simply discarded" the threat of climate change when approving the pipeline.

Democrats will bring their own agendas. But lately, the party has shown more solidarity with science and evidence-based policymaking.

Come January, when the elected candidates assume their positions, science will have a more prominent place at the political table on Capitol Hill. The United States — and indeed, the world — is facing crucial questions about everything from public health and inequality to global warming. Any development that strengthens the voice of evidence, whatever side of the aisle it comes from, is one to support. ■



If you can't build well, then build nothing at all

Scientists must call out — not merely greenwash — infrastructure building that will ruin environments, lives and economies, urges William Laurance.

Esteemed Brazilian scientist Eneas Salati once said that the best thing that could be done for the Amazon was to blow up all the roads. By 2050, Earth could accumulate another 25 million kilometres of paved roads, according to the International Energy Agency — enough to encircle the planet more than 600 times. When a new road penetrates intact forest, it can facilitate illegal deforestation, poaching, fires and land investors bent on encouraging a building boom — factors that are rarely considered in cost-benefit analyses of planned infrastructure projects.

Around nine-tenths of new infrastructure is slated for developing nations, which contain nearly all of the world's tropical and subtropical forests — biologically, the richest real estate on the planet. Yet many 'greening' measures, such as adding rope-bridges or underpasses to help species cross roads, bring relatively trivial benefits, akin to treating cancer with a Band-Aid.

Developing nations unquestionably need better infrastructure, but the benefits of many building proposals are oversold. Even without considering environmental costs, many infrastructure projects risk causing damage to countries' finances, social cohesion and responsible governance. They would not survive a rational cost-benefit analysis that factored in liabilities such as long-term maintenance costs, debt burdens and social well-being.

Instead of roately focusing on mitigating environmental damage, we need to develop global guidelines to assess whether an infrastructure project should even go forward. Establishing a half-dozen simple criteria could red-flag the highest-risk projects, which the global community could research in depth and potentially recommend for cancellation. This task should be adopted in Sharm El-Sheikh, Egypt, this month, when a United Nations meeting of the Convention on Biological Diversity (CBD) will discuss infrastructure and extractive industries, such as mining and petroleum, that spur road development.

If you think conventional environmental-impact assessments are sufficient for countries to make appropriate infrastructure decisions, you are misguided. These evaluations are systematically biased towards project approval, in part because project proponents pay for them or may exert undue influence on government decision-makers.

For example, the environmental-impact assessment for Brazil's 900-kilometre-long BR-319 highway, which is slicing into the heart of Amazonia, concluded that the project would cause no net increase in deforestation. Yet independent analyses suggest that it will provoke dramatic acceleration of forest loss — an extra 5 million to 39 million hectares by mid-century (C. D. Ritter *et al.* *Biol. Conserv.* **206**, 161–168; 2017). Similarly, the provincial government of North Sumatra, Indonesia, approved a hydropower project that would cut across the scarce habitat of the critically endangered Tapanuli orangutan (*Pongo tapanuliensis*), of which there are fewer than 800 individuals still alive.

My colleagues and I found the environmental-impact assessment to be rife with inaccuracies and misinformation, which we reported to Indonesian President Joko Widodo in July. A local non-governmental organization is now challenging the project in a lawsuit.

It is often hard for citizens to access unbiased information about infrastructure projects in their countries. China's Belt and Road Initiative is intended to span some 120 nations and involve at least 7,000 infrastructure and extractive-industry projects. But because the initiative is inscribed in the Communist Party's constitution, it is legally protected from public criticism within China. Bad news about the scheme is blocked by government censors, or simply not translated into Mandarin.

People without these constraints must speak up. Too many scientists are ceding responsibility to overstretched decision-makers and public-interest groups. Some put too much trust in existing regulations and safeguards. Others think that all development is good, or that it's inappropriate to advise a country if you're a foreigner. Some think it's just hopeless. And many simply don't have the stomach for real-world conservation: it's controversial, taxing and stressful. It can also be dangerous: my colleagues and I have faced death threats and lawsuits for speaking out against projects.

But we need to drive home messages that most current assessments won't or can't: that the price of building a road in a flood zone might not include installing proper drainage or rebuilding after inevitable washouts; or that, without an assured funding stream for maintenance, big investments for infrastructure, such as a major paved highway or

hydropower project, can easily be squandered, yet the damage and debt they create remains. Many nations, including Pakistan, Laos, Sri Lanka and some Pacific Island countries, are now veering towards insolvency.

If a project is in a remote area, wilderness or locale prone to flooding, that's a red flag. Another is the likelihood of highly inequitable economic benefits. Many developing nations, including Brazil, Papua New Guinea and Nigeria, have been plagued by such projects. Brazil, for instance, has lost billions of dollars in bad hydro-dam investments.

The CBD needs to set out simple guidelines and priorities to help nations produce smart, sustainable infrastructure. Experts should independently investigate projects pocked with red flags. Every nation has a sovereign right to determine its own development priorities. There is nothing even faintly undemocratic about giving citizens in each nation an opportunity to understand the real risks involved.

Many building projects should be screened out entirely — not just greened up. ■

WE NEED TO
DRIVE HOME
MESSAGES
THAT MOST CURRENT
ASSESSMENTS
WON'T
OR CAN'T.

William Laurance directs the Centre for Tropical Environmental and Sustainability Science at James Cook University in Cairns, Australia. e-mail: bill.laurance@jcu.edu.au

SEVEN DAYS

The news in brief

POLICY

Red-tape reduction

The European Commission has reduced the amount of red tape for scientists applying for grants under its flagship research-funding programme, according to a report released on 6 November by the European Court of Auditors (ECA). The commission had sought to cut bureaucracy in the latest iteration of the €76.4-billion (US\$86-billion) Horizon 2020 framework programme. Measures included centralizing support services and developing a single rule book for participation. These changes reduced the administrative burden on grant applicants, the ECA found, and cut the time between applying for and receiving a grant. But the report also notes areas that need improvement; for example, it says that the commission could do more to help researchers who submit high-quality, but unsuccessful, applications to obtain funding from other sources.

Chemical strategy

The European Commission adopted on 7 November a strategy to crack down on the use of endocrine-disrupting chemicals (EDCs). The chemicals unbalance hormone systems, and evidence suggests that they damage human health and affect wildlife. The 28 European commissioners — one from each member country — approved a long-awaited plan for regulating EDCs, which are found in everyday products as well as in some pesticides and biocides, and are linked to disorders including cancer, obesity and lowered fertility. The plan includes further research, a check on current EDC legislation to pinpoint weaknesses, and the development of improved testing methods.



ANDREW LICHENSTEIN/CORBIS/GETTY

Huge oil-pipeline project blocked

A federal judge in Montana has blocked construction of the Keystone XL pipeline — which would enable transport of oil from the tar sands of Alberta, Canada, to refineries along the Gulf of Mexico. The administration of US President Donald Trump had “simply discarded” the project’s potential impacts on greenhouse-gas emissions, rather than justifying its decision to issue a permit for the pipeline, the judge said

in an 8 November ruling. Former president Barack Obama had rejected the project in 2015 after an analysis suggested that it would increase greenhouse-gas emissions. The state department issued a permit in March 2017 after Trump took office, but environmentalists and Indigenous-rights groups — some of whom had protested against the controversial project (pictured) — challenged the move in federal court.

But critics, including Brussels-based EDC-Free Europe, a coalition of more than 70 environmental, health, women’s and consumer groups, said the plan lacked concrete measures to reduce harmful exposures.

CONSERVATION

Horse culls

Ninety Australian scientists are calling for the repeal of a June 2018 law that protects free-roaming horses in the country’s alpine regions. The Kosciuszko Science Accord demands that the New South Wales government acknowledge the “potentially irreparable

damage” that the horses, which are technically feral, are causing to the iconic Kosciuszko National Park in the state’s southeast. The statement was signed at a conference on the impact of the horses — which are harming plants and fragile ecosystems — held at the Australian Academy of Science’s Shine Dome in Canberra. It also demands that New South Wales, Victoria and the Australian Capital Territory, whose jurisdictions cover the Australian Alps, cooperate to remove the horses through aerial culling, which is banned in New South Wales, or other effective means. Scientists estimate that there

are 7,000–8,000 free-roaming horses in the Australian Alps. See go.nature.com/2rocfrb for more.

EVENTS

Koreas TB deal

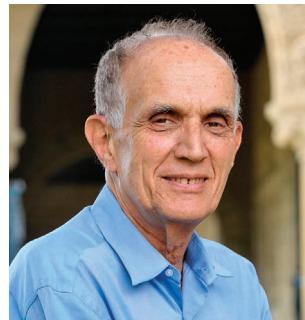
North Korea and South Korea will establish a joint response system for fighting contagious diseases such as tuberculosis (TB), a major public-health threat in North Korea. A pilot programme, set to begin by the end of the year, will see the two nations exchange information on contagious diseases through a liaison office in Kaesong, on the northern side of the border. The agreement

ROD SEARCY is part of a pledge to expand public-health collaboration, laid out in a declaration that South Korean President Moon Jae-in and North Korean leader Kim Jong-un signed in Pyongyang on 19 September. The agreement is a necessary first step in planning for free movement between the Koreas, or for eventual reunification, says Shin Hee Young, a paediatric oncologist and director of the Institute for Health and Unification Studies at Seoul National University. Bacterial diseases, such as TB, rheumatic fever and scarlet fever, that are rampant in North Korea are much less common in South Korea, says Shin. "When these people cross the border without any restrictions, there will be an epidemic of tuberculosis in the South," he says. More than 107,000 cases of TB were reported in North Korea in 2017, resulting in an estimated 16,000 deaths, according to the World Health Organization.

AWARDS

Stats 'rock star'

US statistician Bradley Efron (**pictured**) at Stanford University in California has won the 2018 International Prize in Statistics for pioneering the 'bootstrap' method for measuring the



reliability of small data samples. His work, which dates back to 1977, has given rise to techniques now commonly used across many scientific disciplines. The American Statistical Association — which administers the prize together with four other scientific societies — announced the winner on 12 November. The US\$80,000 prize was first awarded in 2016 and is given out every two years; British statistician David Cox was its first winner. Efron, who is 80, says that he was "thrilled" to receive the prize. Scientists often have to wait many years to get their "round of applause", he says. "It turns out that's okay — it feels great!" Sally Morton, a statistician at Virginia Tech in Blacksburg, says that Efron is "a statistical rock star". "He has inspired generations of statisticians and scientists," she says.

POLITICS

Call for Brexit vote

The parliament of Scotland has become the first UK legislative body to support a public vote on the final terms of any Brexit deal. The United Kingdom is scheduled to leave the European Union on 29 March 2019 and the government is seeking to finalize a withdrawal agreement, which must be approved by the UK parliament and EU member states. On 7 November, Scottish members of parliament voted to pass a motion on the threats that Brexit poses to Scotland's science and research, which included an amendment calling for a 'people's vote'. Withdrawal negotiations have been criticized by many in and out of government as chaotic, and calls for a second public vote have surged in recent weeks. Scientists have repeatedly warned that Brexit could have a catastrophic effect on science and collaboration. Two days later, UK transport minister and former science minister Jo Johnson resigned from the government over the negotiations, and also called for a referendum on the terms of any deal — which he said should include

an explicit option for the United Kingdom to remain in the EU.

POLITICS

Nuclear-power vote

Hundreds of researchers in Taiwan have signed an open letter urging the public to vote to continue the phase-out of nuclear power in an upcoming referendum. Last year, Taiwanese legislators added a clause to the island's electricity act to shut down all nuclear power plants by 2025. But many people disagree with the plan. This October, proponents of nuclear power gathered enough signatures — more than 1.5% of the electorate in Taiwan — to force a referendum that will ask the public to agree to removing the phase-out clause from the act. The vote will be held on 24 November, along with multiple other referendums and local elections. The researchers warn that Taiwan is at high risk of earthquakes and tsunamis — events that can damage nuclear-power stations with devastating effects — and doesn't yet have a feasible long-term solution for dealing with the radioactive waste. The waste is currently stored at the power stations or on Orchid Island off the east coast.

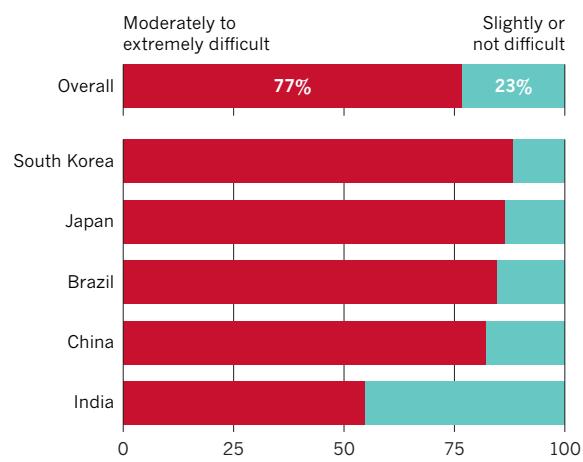
TREND WATCH

More than two-thirds of researchers find it difficult to prepare manuscripts and respond to peer-review comments, finds a survey of nearly 7,000 researchers in 100 countries, released on 9 October. The issues might stem from language barriers, suggests the report. Some 70% of respondents were based in emerging scientific powerhouses: Brazil, China, India, Japan and South Korea. Only 11% had English as a first language, and 45% said that they found it difficult to write in English. The poll was carried out between December 2016 and January 2018 by Editage, a company based

in Philadelphia, Pennsylvania, that offers language-editing and publication support. Two-thirds of respondents who had authored papers felt that journal guidelines in general were unclear, incomplete or both, and three out of four said that preparing manuscripts in English was the most challenging part of publishing. The publishing industry needs to consider how to eliminate or minimize the extra burden on authors whose first language isn't English, the report says. Otherwise, journals risk missing out on research because authors might choose to submit to regional-language publications.

ENGLISH-LANGUAGE BARRIER

Researchers in major scientific nations often struggle to navigate publishing requirements in English-language journals. Many say they find it difficult to write and to prepare their manuscripts in English.



NEWS IN FOCUS

POLITICS Science candidates win in US midterm elections **p.302**

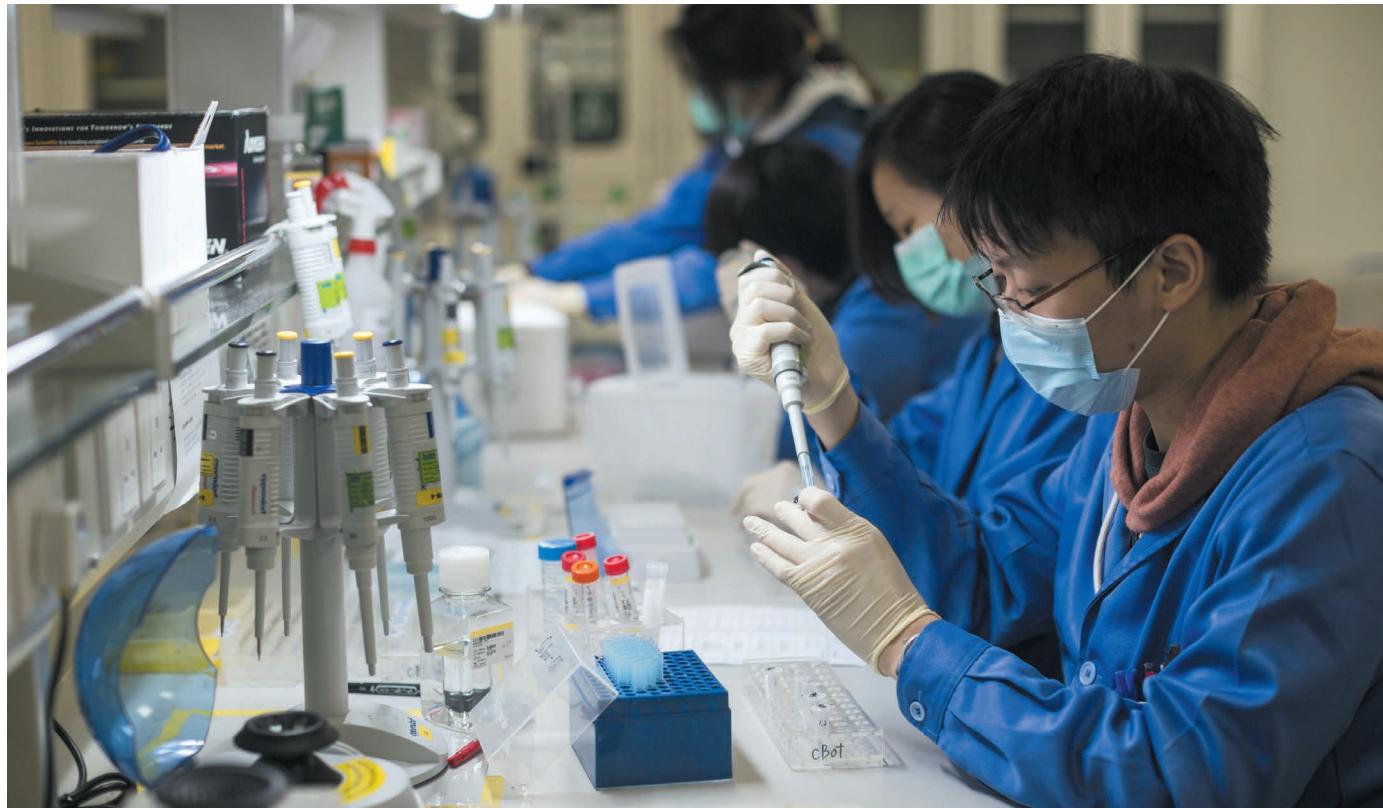


GENETICS Trove of ancient DNA rewrites history of South America **p.303**

NEUROSCIENCE Body's secrets revealed in see-through mice **p.305**

MEDICINE How participants are changing the way clinical trials are run **p.312**

JEROME FAVER/EP/SHUTTERSTOCK



In China, samples of human DNA cannot be shared between companies or institutions without permission from the government.

POLICY

China cracks down on genetics breaches

Biomedical companies have been punished for sharing DNA data without permission.

BY DAVID CYRANOSKI

China's enormous population is a genetics goldmine. But the government, wary that these data could be exploited for profit, has been cracking down on researchers and companies that violate rules on sharing its citizens' genetic material and information. Some scientists fear that this closer attention is creating hurdles for international collaborations.

Last month, for the first time, the ministry

of science named and shamed companies that have broken the sharing regulations that the government introduced in 1998. Five companies and one research hospital were rebuked for transferring human DNA samples or genetic data to other entities in China or outside the country, without permission from the ministry's human genetic resources office. It is not clear why the ministry released details of the breaches now — some as recent as this year, others a few years old.

Global pharmaceutical giant AstraZeneca,

which has a research centre in Shanghai, was caught earlier this year transferring samples — used to create diagnostic tests for predisposition to breast cancer — to two smaller Chinese companies, Amoy Diagnostics in Xiamen and Kunhao Ruicheng in Beijing. AstraZeneca was authorized to collect the samples, but the company says it did not know that it needed permission to transfer the material to another party in China.

The regulations require government authorization for anyone who wants to transfer ▶

► human DNA samples or share genetic data. Permission is also required to publish these data in international journals.

The ministry says genomics giant BGI in Shenzhen and Shanghai's Huashan Hospital were also caught breaking the regulations, after they put genetic information online without approval. The data were part of a large international study on the genetics of depression, which was published in *Nature* in 2015 (CONVERGE consortium. *Nature* 523, 588–591; 2015). The paper was based on anonymized sequence data from more than 10,000 Chinese women, which BGI acknowledges it did not have permission to publish in the paper's supplementary material.

A spokesperson for the company says it has destroyed the data, as requested by the ministry. They say the company has also requested *Nature* remove the article from its website. It remains online. A spokesperson for *Nature* would not comment on the matter. (*Nature*'s news team is editorially independent of its journal team.)

Scientists and policy experts are worried that the government crackdown might deter researchers from sharing genetic data collected in China. "At a time when transparency, open access and sharing are high priorities, enforcing the 1998 rules obviously seems to be going in the opposite direction," says Nicholas Steneck, who studies research integrity at the University of Michigan in Ann Arbor.

Many countries control how their citizens' genetic material and data can be collected

and shared, mainly to protect people's privacy and ensure that samples are gathered with informed consent. China's rules are also meant to ensure that the country reaps some of the profits from patented discoveries.

But scientists say that complying with the rules is creating obstacles. An international collaboration investigating genetic samples from more than 140,000 pregnant Chinese women had to send a data-analysis expert to China because the data could not leave the

"If applying for permission is onerous or time-consuming, this will have a detrimental effect."

country, says group member Anders Albrechtsen, a geneticist at the University of Copenhagen.

The group — which included researchers from BGI — did not try to get approval to publish the anonymized genetic data. Instead, in a paper published in *Cell* in October, it included a disclaimer saying that the authors will provide only summary statistics to other researchers (S. Liu *et al.* *Cell* 175, 347–359; 2018). The president of BGI Research, Xu Xun, says the team feared that it would have taken too much time and effort to get permission to share the raw sequence data. He also thinks that sharing population-level statistics is sufficient.

Geneticist Paul Flicek of the Wellcome Sanger Institute in Hinxton, UK, thinks it

is reasonable for governments to require approval to share genetic information, but that "if the process of applying for permission is onerous or time consuming, this will have a detrimental effect on data sharing".

If China continues to enforce its regulations, genetics research in the country could become isolated from international groups, says Arcadi Navarro, a geneticist at Pompeu Fabra University in Barcelona, Spain.

A spokesperson for *Cell* says that the journal requires that the data behind publications be made available, but its policy acknowledges the need to respect the regulations and guidelines of review boards and national bodies, as well as laws on patient privacy and personal data.

China's science ministry did not respond to *Nature*'s questions about whether its restrictions impede research.

In its announcement, the ministry did say that, as punishment for their breaches, BGI, AstraZeneca and Huashan Hospital had been banned from participating in international collaborations that use human genetic resources until they passed a data-privacy examination. BGI says it passed this in 2017. AstraZeneca says it is working towards its reassessment now. *Nature*'s attempts to contact the hospital were unsuccessful.

Both BGI and AstraZeneca say that they accept the government's penalties and support the country's attempts to protect the genetic resources of its citizens. ■

POLITICS

Scientists win in US midterm elections

Trump administration's controversial science and environment policies could come under extra scrutiny as Democrats gain in Congress.

BY JANE J. LEE, AMY MAXMEN, JEREMY REHM & JEFF TOLLEFSON

The results of the political experiment are in. At least 12 candidates with backgrounds in science, technology, engineering or medicine were elected to the US House of Representatives on 6 November — including several who had never before run for political office.

They include Elaine Luria, a US Navy veteran and nuclear engineer in Virginia, and Chrissy Houlahan, a former business executive with a degree in engineering, in Pennsylvania. Illinois saw wins by registered nurse Lauren Underwood, a former senior adviser

to the Department of Health and Human Services, and clean-energy entrepreneur Sean Casten, who has degrees in engineering and biochemistry.

The four — all Democrats — are among roughly 50 candidates with science backgrounds who ran for the House in 2018, sparked in part by opposition to President Donald Trump. Fewer than half of these novice politicians made it past the primaries to the general election, but many science advocates are already looking to the next campaign cycle.

"I'm feeling good," says Representative Bill Foster (Democrat, Illinois), a physicist who has pushed to increase the number of scientists in elected office. Foster, the only current member

of Congress with a science PhD, is excited about wins at the state and local levels by candidates with backgrounds in science, technology, engineering or medicine (STEM). "We'll have a much deeper bench among STEM candidates in future races for Congress," he says.

The advocacy group 314 Action, which sprang up after the 2016 election to help scientists run for office, says that 8 of the 22 candidates it endorsed for the House or Senate ultimately won. The group in Washington DC also backed about 50 candidates in state races, and 31 won.

"It's certainly exceeded our expectations of what we would be able to do this year," says Shaughnessy Naughton, 314 Action's

TOM WILLIAMS/CO-ROLL CALL/GETTY
president. She says that the group spent US\$2 million during this election cycle on items such as ads and voter-registration drives, and contributed another \$250,000 to various candidates' campaigns.

That wave of interest is "indicative of people's desire to get involved, and a recognition that it's no longer okay to sit on the sidelines", says Benjamin Corb, director of public affairs at the American Society for Biochemistry and Molecular Biology in Rockville, Maryland.

The victories for science candidates came as Democrats regained a majority of seats in the House, taking the chamber back from Republicans — who still control the Senate and the White House. Recapturing the House is "no small feat", says Elizabeth Gore, senior vice-president for political affairs at the Environmental Defense Fund, an advocacy group in New York City. "It is going to change the dialogue in Washington, and will certainly change the dynamic around science and the environment."

A CHANGING CLIMATE

One of the most dramatic transitions will involve the House Committee on Science, Space and Technology. Representative Eddie Bernice Johnson, a Texas Democrat and vocal critic of the Trump administration, is likely to take the helm from retiring Representative Lamar Smith (Republican, Texas). As chair, Smith has repeatedly questioned the science behind climate change, sought to pare back the National Science Foundation's research portfolio and launched dozens of probes into alleged wrongdoing by individual scientists and US government science agencies.

By contrast, Johnson released a list of policy priorities on 6 November that includes fighting climate change — "starting with acknowledging it is real" — and making the science panel "a place where science is respected".

Smith is not the only Republican with a strong



Eddie Bernice Johnson (left) is in line to become the next leader of the House science committee.

interest in science who will exit Congress at the end of year. Voters rejected a bid for re-election by Representative John Culberson of Texas, a space enthusiast who leads the House spending panel that oversees NASA, the National Science Foundation and the National Oceanic and Atmospheric Administration. Culberson's stalwart support for a NASA mission to Jupiter's moon Europa became a campaign issue after his opponent accused him of favouring pet projects and neglecting local issues in his district near Houston.

Culberson is "probably the strongest supporter of planetary science, maybe in history", says Casey Dreier, senior policy adviser at the Planetary Society in Pasadena, California. "It was so neat to see someone in Congress who had a personal passion for the search for extraterrestrial life."

Holding even a slim margin in the House will give Democrats the power to investigate

the Trump administration's policies. Gore says that this is likely to translate into congressional hearings that probe the administration's efforts to roll back a variety of climate and environmental regulations, and explore whether they are justified by the available science.

"Some of the oversight that we will see in a Democratic House will be focused on re-establishing scientific integrity and highlighting the failure of the Trump administration to use scientifically based information for policy-making," Gore adds.

Others worry that with Democrats taking the House and Republicans solidifying their majority in the Senate, political gridlock will worsen in the coming years. "The polarization in the Congress has increased," says Robert Stavins, an environmental economist at Harvard University in Boston, Massachusetts. "What was left of moderate Republicans — those are the people who systematically lost to Democrats." ■

ANCIENT GENOMICS

Migration to Americas traced

Genomes show that the Americas' earliest settlers moved far and fast across the continent.

BY EWEN CALLAWAY

Ancient genomics is finally beginning to tell the history of the Americas — and it's looking messy.

Genomes from dozens of ancient inhabitants of North and South America, who lived as much as 11,000 years ago, suggest that the populations moved fast and frequently. The findings, published on 8 November^{1,2}, indicate that North America was populated widely over a few hundred years, and South

America within 1,000–2,000 years by related groups. Later migrations on and between the continents connected populations living as far apart as California and the Andes.

"These early populations are really blasting across the continent," says David Meltzer, an archaeologist at Southern Methodist University in Dallas, Texas, who co-led one study².

The studies also suggest that the prehistory of the Americas — the last major land mass to be settled — was just as convoluted as that of other parts of the world.

"I think this series of papers will be remembered as the first glimpse of the real complexity of these multiple peopling events," says Ben Potter, an archaeologist at the University of Alaska Fairbanks. "It's awesome."

For decades, the peopling of the Americas was painted in broad brushstrokes, using data from archaeological finds and DNA from modern humans. Scientists discerned that groups crossed the Bering land bridge from Siberia into present-day Alaska, and then moved steadily south as the last ice age ended. Humans ▶



An arrowhead that belonged to people associated with the Clovis culture, early settlers in the Americas.

▶ carrying artefacts, such as sophisticated projectile points, from a culture known as Clovis began to populate the interior of North America about 13,000 years ago. For decades, scientists thought that people associated with this culture were the continents' first inhabitants.

But the discovery of 'pre-Clovis' settlements — including a nearly 15,000-year-old site at the southern tip of Chile — pointed to an even earlier wave of migration to the Americas.

The first ancient-DNA studies from the region, appearing in 2014, began to add detail to this picture. The genome of a baby boy who was buried roughly 12,700 years ago in Montana alongside Clovis artefacts³, and genomes from other ancient individuals⁴, hinted at two early populations of Native Americans.

The Montana baby, known as the Anzick boy, belonged to a population known as the Southern Native Americans, who are most closely related to present-day Indigenous populations from South America. They split from Northern Native Americans, who are genetically closer to many contemporary groups in eastern North America, around 14,600–17,500 years ago. And

the common ancestor of those two groups split from East Asians some 25,000 years ago, as scientists established earlier this year by sequencing the genome of 11,500-year-old human remains from Alaska⁵.

But this timeline was based on just a few ancient genomes from the Americas, and scientists expected further data to paint a more detailed, complex picture of the continents' history, as well as reveal later migrations there.

SAME GENES, FAR APART

The two latest studies include genome data from 64 ancient Americans, and provide the first detailed look at the ancient inhabitants of Central and South America and their early movements into the region.

To chart these migrations, Meltzer and his colleague Eske Willerslev, a palaeogeneticist at the Natural History Museum of Denmark in Copenhagen, compared genetic data from the Anzick boy with those from 10,700-year-old remains in a Nevada cave and 10,400-year-old remains from southeastern Brazil.

The genomes were remarkably similar,

despite the great geographical distances between them, Willerslev says, pointing to a rapid population expansion from Alaska. "As soon as they get south of the continental ice caps, they're exploding and occupying the land," he says.

An independent team led by David Reich, a population geneticist at Harvard Medical School in Boston, Massachusetts, also found evidence¹ for a rapid expansion into South America, through analysing 49 ancient genomes from Central and South Americans.

Both teams documented multiple later human migrations into South America. Reich's group found, for instance, that the genetic signal of the earliest inhabitants — closely related to the Anzick boy — had largely vanished from later South Americans, suggesting that different groups had by then moved in from the north.

Potter says that the main conclusions of the two papers are broadly consistent. "Complex and realistic are the two adjectives I would use," he says.

Even with dozens more newly discovered ancient genomes from the Americas, important aspects of the region's population history are probably still missing, says Reich. "There are many dots that are not filled in," he says. "I think as these studies scratch the surface, they make things more, rather than less, complicated."

Jennifer Raff, an anthropological geneticist at the University of Kansas in Lawrence, says that the emerging picture of the Americas is less a revision of the earlier models and more an elaboration. "It's not that everything we know is getting overturned. We're just filling in details," she says. ■

- Posth, C. *et al.* *Cell* <https://doi.org/10.1016/j.cell.2018.10.027> (2018).
- Moreno-Mayar, J. V. *et al.* *Science* <https://doi.org/10.1126/science.aav2621> (2018).
- Rasmussen, M. *et al.* *Nature* **506**, 225–229 (2014).
- Rasmussen, M. *et al.* *Nature* **523**, 455–458 (2015).
- Moreno-Mayar, J. V. *et al.* *Nature* **553**, 203–207 (2018).

enough people, and that its scope may have been too narrow. Its members, who say they are among 12 people who contributed evidence to the April complaint that prompted the probe, also question the investigation's transparency.

Their concerns "cast doubt as to whether the investigation was conducted in a manner that was as effective as it could be, given the seriousness of the allegations", they say in a statement seen by *Nature*. On 2 November, Serena Nik-Zainal, a clinical scientist who now works at the University of Cambridge, sent the statement to Genome Research Limited (GRL), which oversees the Sanger and commissioned the investigation from barrister Thomas Kibling. "We firmly believe sufficient evidence was not unearthed to make an appropriate judgement," says the statement.

David Willetts, chair of the board of GRL, told *Nature* that the investigation was independent

INSTITUTIONS

Sanger whistle-blowers dispute inquiry findings

Leading genomics institute stands by conclusions of an investigation that clears its management of bullying.

BY HOLLY ELSE

Six current and former employees are calling for the Wellcome Sanger Institute in Hinxton, UK — one of the world's top genomics centres — to reopen an investigation that last month cleared its management of

bullying, gender discrimination and misuse of grant money.

The group raises concerns about the process of the investigation and questions the decision to clear senior management at the institute of the allegations. Among other things, the group says that the investigation did not interview

and detailed, and that the organization does not plan to review the findings. "We believe Mr Kibling carried out a thorough and independent investigation as he was tasked to do," he says.

The Sanger employs almost 1,000 scientists and other skilled professionals, and played a key part in the Human Genome Project, which concluded in 2003.

On 30 October, GRL released a redacted executive summary of Kibling's investigation report. The summary said that the investigation considered "various whistleblowing concerns" in a document submitted by one staff member that alleged that the institute and its director, the geneticist Mike Stratton, had committed gender discrimination, wrongful exploitation of scientific work for commercial purposes and misuse of grant monies. The summary also says that the investigation considered an allegation that Stratton had bullied someone. And it says that Kibling, of Matrix Chambers in London, cleared Stratton and the Sanger's management of all these accusations.

The authors of the 2 November statement are Nik-Zainal; Inés Barroso, a human geneticist who has worked at the Sanger since 2002 and who says she wrote the initial whistle-blowing complaint; Jyoti Choudhary, a proteomicist now at the Institute of Cancer Research in London; and three people, including a former member of the senior management team, who wish to remain anonymous to protect their careers.

Their statement questions the level of information that the investigation considered. It also questions the investigation's finding that there is no evidence for some allegations, and suggests that this might be because crucial evidence fell outside the scope of the investigation.

In his summary, Kibling notes that he was not required to determine the merits of any individual's grievance "which are not in the nature of a whistleblowing complaint or advanced by others" — and that such grievances are to be dealt with in a separate process.

Kibling told *Nature* that he stands by his investigation, and it was his "judgement call" to

decide who would assist him and therefore who to interview. "The investigation needs to be proportionate and focused on the whistle-blowing complaint made and not the individual grievances that some of those I spoke to harboured," he says. He adds that he believes that he spoke to those who had a valuable contribution to make and were necessary for the investigation.

The investigation did identify failings in how people have been managed at the Sanger, and a lack of diversity at senior levels of the organization. The 2 November statement acknowledges these findings, but the authors still say that they are "disappointed by the investigation process".

They call on the Wellcome Trust in London, which owns the Sanger, "to reconsider whether the principles of this investigation lived up to its own standards".

Wellcome says that it is "satisfied with the investigation that has been carried out", and has no plans to reopen the probe.

Stratton did not respond to *Nature*'s request for comment. ■

NEUROSCIENCE

'Invisible' mice reveal anatomical secrets

Technique that turns dead rodents clear uncovers surprising details about injury response.

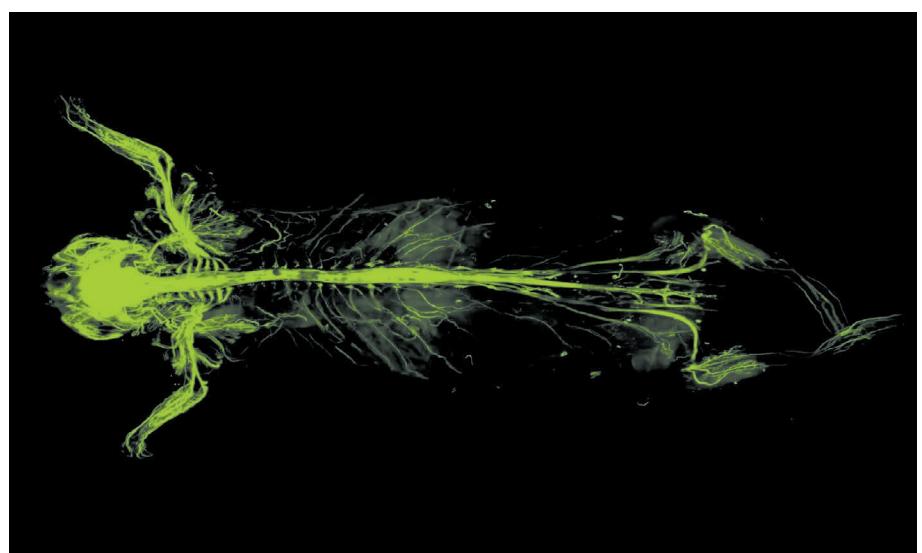
BY SARA REARDON

A new technique that makes dead mice transparent and hard like plastic is giving researchers an unprecedented view of how different types of cell interact in the body. Scientists can pinpoint specific tissues while scanning an animal's entire body.

The approach, called vDISCO, has already revealed surprising structural connections between organs, including hints about the extent to which brain injuries affect the immune system and nerves in other parts of the body. That could lead to better treatments for traumatic brain injury or stroke.

Methods that turn entire organs clear have become popular in the past few years, because they allow scientists to study delicate internal structures without disturbing them. But removing organs from an animal's body for analysis can make it harder to see the full effects of an injury or disease. And if scientists use older methods to make an entire mouse transparent, it can be difficult to ensure that the fluorescent markers used to label cells reach the deepest parts of an organ.

The vDISCO technique overcomes many of these problems. By making the dead mice



The nervous system of a mouse treated using the vDISCO technique glows green.

rigid and see-through, it can preserve their bodies for years, down to the structure of individual cells, says Ali Ertürk, a neuroscientist at Ludwig Maximilian University of Munich in Germany, who led the team that developed vDISCO. He presented the work this month at

a meeting of the Society for Neuroscience in San Diego, California.

The process begins by soaking a mouse's body in organic solvents to strip it of fats and pigments. This preserves the structure of cells, even as the mouse shrinks by up to 60% (ref. 1). ▶

► To explore the transparent mice, Ertürk's team developed a way to home in on specific cell types, such as neurons or cancer cells. The scientists turned to 'nanobodies': antibodies that are found only in llamas, camels and alpacas, and are one-tenth the size of antibodies in other species.

Nanobodies can be engineered to stick to specific proteins that are found only in one type of cell — while carrying fluorescent green markers to label the chosen cells. And because nanobodies are so small, they can easily pass through tiny blood vessels and into organs.

When the researchers pumped nanobodies into the circulatory systems of dead mice, which carried the molecules throughout the body, they could see individual cells glowing bright green under a microscope.

The technique is the first to make whole animals truly transparent, says Kwanghun Chung, a medical engineer at the Massachusetts Institute of Technology in Cambridge. "I think it's a fantastic technology," he says.

The first experiments with vDISCO have yielded surprising discoveries. One involves mysterious vessels that run between the skull and the brain, which

were discovered only in 2015 (ref 2). When a team led by Ruiyao Cai, a

neuroscientist in Ertürk's lab, used nanobodies to light up lymphatic vessels in a mouse treated with vDISCO, the vessels in the head glowed green — confirming scientists' suspicions that the structures are part of the system that transports lymph.

Cai and Ertürk also used vDISCO to test how severe injuries to the brain and spinal cord affect cells elsewhere in the body. Labelling neurons showed that nerves in a mouse's torso degraded after the animal suffered a traumatic brain injury, even though the nerve cells were far from the injury site. In another case, the scientists spotted immune cells that had rushed to the site of a spinal-cord injury days before a mouse died — and, unexpectedly, into surrounding muscle and lymphatic vessels³.

The combination of vDISCO and nanobodies is "kind of the direction for the future", says Hiroki Ueda, a biologist at the University of Tokyo.

Ertürk next plans to use vDISCO to trace how viruses, cancer cells and other invaders spread throughout the body. His group is also designing machine-learning approaches to count and assess labelled cells without introducing bias or human error. ■

- Pan, C. et al. *Nature Meth.* **13**, 859–867 (2016).
- Louveau, A. et al. *Nature* **523**, 337–341 (2015).
- Cai, R. et al. Preprint at bioRxiv <https://doi.org/10.1101/374785> (2018).



The destruction of olive trees infected with a bacterium has caused controversy in Italy.

ITALY

Deadly olive-tree disease spreads

Measures meant to stop bacterium's expansion across Italy have been delayed multiple times.

BY ALISON ABBOTT

A vicious bacterium devastating Italy's valuable olive groves is still spreading years after it was identified, because of opposition to measures meant to contain the pathogen.

After months of inaction, authorities in the Puglia region have now resumed efforts to track the spread of *Xylella fastidiosa*, which causes a disease called olive quick-decline syndrome (OQDS) that cannot be cured or eradicated.

But scientists say the delays in implementing disease-containment measures — Italy declared a state of emergency over the outbreak in early 2015 — have added to the growing risk that the infection will spread out of the Puglian peninsula, which lies within the heel of Italy's 'boot', and towards olive groves in the country's main landmass.

Quarantine efforts — which environmentalists and farmers have frequently opposed

— stopped again most recently in May. In the same month, the European Commission extended the 'certified infection zone' where the disease is present by 20 kilometres.

The delays have been a problem, says plant pathologist Maria Saponari of the Institute for Sustainable Plant Protection in Bari, Puglia's capital: "The later you detect an infection, the later you can start all the containment actions that are needed."

The budget now allocated by the Puglian government to begin tracking the bacterium again — €1.8 million (US\$2 million) — also falls short of what is needed to implement the full set of containment measures agreed to by the Italian government and the European Commission four years ago.

Italy could now face legal consequences for its inaction, after the European Commission made good in May on its longstanding threat to refer the nation to the European Court of Justice for violating its quarantine regulations. If found guilty, Italy could, for example, lose

FABIO SERINO/ROPI/VIA ZUMA

access to important agricultural subsidies.

The bacterium had never been seen in Europe until 2013, when it was identified in southern Puglia. The outbreak was immediately subjected to stringent European Union quarantine regulations, which were agreed with the Italian government.

The original containment plan dictated that infected trees be uprooted and destroyed, as well as the apparently healthy trees surrounding them. It also required the application of insecticides to control spittlebugs, which transfer the bacteria between trees.

STRONG OPPOSITION

But environmentalists and some farmers have objected to these practices — and some have claimed that the containment measures were based on false science.

Politicians have wavered over whom to please, and protests and court cases have often stopped the activities. Some trees identified as infected through monitoring activities earlier this year remain standing. And in spring, mayors of eight communities in Puglia publicly declared that they would not comply with the insecticide requirement.

The area affected by the bacterium has expanded steadily since 2013. The European Commission's May update on the situation designated the whole of south and central

Puglia as an infected zone, and a region to its north as a buffer zone that must be also carefully monitored for new cases of the disease.

Italy's agriculture minister, Gian Marco Centinaio, promised in July to propose a full containment plan within a few months, but has not done so, despite public nudges from EU and Puglian politicians. The agriculture ministry did not respond to *Nature*'s request for comment.

"The later you detect an infection, the later you can start the containment actions that are needed."

Last month, an association of olive growers called Coldiretti Puglia sent the government a list of proposals for the containment plan. The group wants special rules to be put in place to stop regional administrative courts from blocking containment measures. It also suggested tapping into national disaster funds that could be used to support the development of new *Xylella*-resistant olive trees and expand the monitoring programme.

GOVERNMENT HANG-UP

The long-running affair and its handling are now being dissected in Italy's parliament. In June, some parliamentarians formally

deposited documents at the Senate, one of Italy's two houses of parliament, which challenged the scientific evidence on which *Xylella* management plans have been based. It also called for a Senate inquiry into whether scientists have misled the public. An independent analysis commissioned by the national science academy, the Accademia dei Lincei, repudiated these claims the following month in an article (see go.nature.com/2t5x1ai). The Senate has not yet acted on the call for an inquiry.

The proposal has fortunately not been carried forward, says Michele Morgante, a plant geneticist at the University of Udine in Italy. Still, he says, it is disturbing that the anti-science activities have received attention at such a high political level.

Meanwhile, in a series of hearings launched independently in September by the Chamber of Deputies — Italy's second house — parliamentarians have interviewed scientists, olive producers and other stakeholders about the *Xylella* outbreak and what could be done about it; hearings are scheduled to continue into next month.

Morgante welcomes the hearings — but says they have come too late: "It is good that parliament [the Chamber of Deputies] finally wants to listen to scientists, but they should have paid attention much earlier when it would have been easier to control," he says. ■

AGAINST ALL ODDS: SCIENCE IN THE PALESTINIAN TERRITORIES

BY ALISON ABBOTT

Travel restrictions and paltry funding hamper researchers, who are trying to build a scientific base.

Dalal Saeed is very clear about what she wants from life: an academic research career in geochemistry. But there are no PhD programmes in the natural sciences in the Palestinian territories. So, every workday, she travels from her West Bank village, across the concrete wall that divides her homeland from Israel to the Hebrew University of Jerusalem, where she has just started her doctoral work.

It's barely 10 kilometres, but the first few times she made the journey through the nearest, traffic-choked checkpoint, it took her more than three hours. She soon learnt to drive farther along the wall to a quieter checkpoint, and halved her travel time.

After decades of conflict, many embittered Palestinians from the occupied territories boycott any form of economic or cultural activities with Israel — including research. But higher education is an exception. "It is an individual decision," says Saeed, who has so far found long travel times her only challenge. Her Israeli co-supervisor, geochemist Boaz Lazar, helped to organize a multi-entry permit for her to enter Israel during the daytime, a scholarship courtesy of the University of Haifa and a project to measure heavy-metal isotopes in and around the Dead Sea, which borders Israel, the West Bank and Jordan.

Saeed has a strong affinity with the Dead Sea, having grown up near its shores and studied it for her master's project. The opportunity to do a PhD "is a dream come true", she says. Once qualified, she'd like to move on to a Palestinian university, but is open to wherever life may take her. Who knows what can happen in three or four years, she asks.

That's a question that many scientists in the territories ask themselves.



They view their predicament as unjust and unstable. But whether Israeli occupation ends in the creation of an independent Palestinian state or in the absorption of the territories into the state of Israel, they say their future depends on having a strong intellectual base and the capacity to carry out research.

Anger in the Palestinian territories festers as conditions for its residents grow increasingly desperate. Travel and imports are heavily controlled by the Israeli military, so researchers find it difficult — and sometimes impossible — to get to international conferences or labs and to access research materials. This year, political tensions have ratcheted up — particularly in July, when the Israeli government passed the Jewish Nation-State bill, which conferred lower citizenship status on non-Jews, including the 20% or so of citizens who are Arabs. Movement restrictions seem to have tightened, too. And Palestinian authorities have devoted little financial support to building up science in the territories.

On the other side of the separation barrier, many — although not all — Israeli scientists oppose how people in the Palestinian territories are being treated. And there have been sharp disagreements in the academic community, especially over the legal status of a university in a Jewish settlement in the territories.

Despite the problems, some researchers and educators are doing what they can to develop a functioning scientific community there. They are setting up research groups for young scientists and taking advantages of opportunities to train overseas and get research grants from foreign governments. "We have so many challenges," says inorganic chemist Abdullatif AbuHijleh, president of Birzeit University near Ramallah.



HEIDI LEVINE FOR NATURE

“But we work hard, we do research and we make progress.”

Israel took control of the Palestinian territories — the regions of East Jerusalem, the West Bank and Gaza — after the 1967 Six-Day War between Israel and its Arab neighbours. (The United Nations and much of the world refer to these as the occupied territories, whereas Israel calls them disputed territories.) During the First Intifada, or Palestinian uprising, in the 1980s, Israeli forces frequently closed Palestinian universities because of suspicions that they might be nurturing attacks on Israelis. The signing of the Oslo Accords in 1993 led to serious peace talks, with a goal of ending the occupation. The Palestinian Authority was established as a governing administration and momentum seemed to be moving towards an independent Palestinian state. But the talks failed and violence resumed in 2000 with the Second Intifada.

Since then, Israeli settlements have expanded into the Palestinian territories and Israel has constructed the separation barrier, which loops protectively around the new settlements. In the past year, tensions have spiked as Palestinians have thrown firebombs and explosives across Gaza’s border fence and launched mortar bombs into Israel. Israeli troops have responded with tear gas, live ammunition and air strikes — an escalation to levels of violence not seen since 2014.

Science has not been a high priority for the Palestinian Authority, which has rarely allocated money to research. But education minister Sabri Saidam’s 2017–22 strategic plan strives to develop research capacity. Last year, his ministry announced a modest, 20-million-shekel (US\$5.5-million) research fund — the first such science budget in 5 years — to be shared between the 14 universities in the territories and the 2,200 full, associate and assistant professors employed there. “It is a signal” of support for science, says Isam Ishaq, assistant president for research at Al-Quds University, one of the territories’ leading universities, in East Jerusalem. Still, he says, it is a constant struggle for universities to find the money to cover running costs for science, and to keep any big equipment in working order.

(The situation is even worse for scientists in Gaza, a strip of coastal land separated from the West Bank by Israeli territories. Electricity is limited to a few hours a day, ruling out most forms of experimental research.)

In the West Bank and East Jerusalem, the Palestinian Authority’s lack of focus on science “is a major gap”, says Sari Nusseibeh, a philosopher at Al-Quds University and a leading academic in the region. Nusseibeh was president of his university in the optimistic 1990s, when he strongly encouraged the development of research — as well as academic cooperation with Israel, a powerhouse for world-class research. Back then he reasoned that if the Palestinian territories were to become an independent state, they would need a strong base in research — not least because they have few natural resources. “As Palestinians, our only resource for self-improvement is ourselves as human beings, and the more initiative we have, the better.”

After the violence of the Second Intifada worsened prospects of an end to the occupation, the idea of a non-violent boycott of Israel gained ground in the territories. Many academics outside the Middle East today boycott cooperation with researchers in Israel, although this movement is much stronger in the humanities and social sciences than in natural sciences, where it has had minimal impact, say Israeli scientists.

Inside the territories, Nusseibeh says it was never clear to him if the mandate of the Boycott, Divestment and Sanctions movement should apply to Palestinians wishing to forge academic links with Israel. “But local pressure increasingly grew to put an end to scientific cooperation,” he says.

Palestinian universities respect the boycott of Israel at the institutional level, but don’t ban individual academics from working together. A few brave souls risk the wrath of public opinion by doing so, although they tend not to broadcast it loudly. This is evident in the fact that applications

Left: Abdul-Rahman Sawalma extracts DNA from a blood sample at Al-Quds University in the West Bank. **Right:** Palestinians cross the Qalandiya military checkpoint to enter Jerusalem in the early morning.

continue to come into the German Research Foundation (DFG) for its trilateral programme, in which Germans and Israelis collaborate with Palestinian scientists, according to a DFG spokesperson.

OUTSIDE FUNDING

More than €71 million (US\$81 million) has been distributed in these projects since the DFG programme started in 1995. Other international programmes have sprung up specifically to help Palestinians. These include the Palestinian–German Science Bridge, a €12.5-million, 5-year programme supported by the German science ministry to help postgraduate students get training in Germany. And last year, the Quebec region of Canada launched a 4-year, million-dollar programme to bring 60 researchers from the territories to the province for 3–5-month research missions.

Palestinians rely heavily on such international programmes, modest as they are, as well as on European Union programmes, which have in the past decade or so transferred nearly €3 million to scientists in the territories who are participating in collaborative projects with people in EU countries.

Although their access to funding is limited, scientists there say that an even bigger impediment to carrying out research is the Israeli occupation. One major problem is a lack of free movement: most people in the West Bank need a permit to enter Israel, and their applications often involve major delays or rejections. Scientists throughout the territories also have trouble importing reagents and equipment because that requires approval from Israeli security channels. Some basic items, such as the fertilizer ammonium nitrate or simple acids, are listed as 'dual use' and are banned, for their alleged potential to be used in weapons. What's more, the isolation of the region has meant that the research community has remained too small and underfunded to be able to offer PhD programmes. Still, the challenges have not killed ambition.

Some Palestinian scientists working in other countries avoid politics and instead organize practical support that they hope will help create human capital for future high-level research back home. Nanotechnologist Mukhles Sowwan keeps a close eye on his former lab at Al-Quds from his position at the Okinawa Institute of Science and Technology in Japan. He helps guide master's students at Al-Quds, assisting the brightest in

is late, has missed a small deadline or has written a careless e-mail. He also discourages any discussion of politics in his lab, which he says can divert attention and cause problems. "Our focus has to be strictly on science and professionalism," he says. No one objects. The atmosphere is eager, expectant.

Only 33 years old himself, Herzallah sees PNI members as the seeds of a future international-level research hub in the territories, and he instructs some of his group in team-leading skills. Abdul-Rahman Sawalma, who already has a medical degree and intends to move to Germany next year to complete a master's and then a PhD, is one of those getting leadership training. Once in Germany, Sawalma will regularly Skype his group at the PNI, just as Herzallah does, to share the knowledge he gains while supervising related research activity there. After his PhD, he wants to do a neurology medical specialization in Germany, but his firm intention is to then return home. "It feels great to be building something pioneering in Palestine," he says.

Many others struggling to get research done in the territories find the topic of politics hard to avoid. The permit and visa issues are a constant reminder. Palestinians who live in the West Bank are not allowed to fly to other countries from Israel's airports without a special permit that they say is practically impossible to get. Instead, they usually first have to travel overland to Jordan, which can add an extra day to a trip.

TRAVEL TROUBLES

Some West Bankers require permits even to move within the West Bank. Young male scientists say that they, in particular, are often stopped by the military for inspections between checkpoints. They asked for their names to be withheld because they feared getting on Israeli security lists. "Every day, the situation gets worse," says one, a sentiment echoed many times. Polymer chemist and Al-Quds vice-president for science and society Hasan Dweik says bluntly: "We are in a big prison."

Foreigners need a visa from Israel to enter the occupied territories, and Palestinian universities have reported a sudden increase in the number of faculty members who have had visa problems. A survey carried out by the Palestinian Ministry of Education found that in the past two academic years, more than half of the 64 foreign faculty members in Palestinian universities have had visas denied or delayed without expla-

"AS PALESTINIANS, OUR ONLY RESOURCE FOR SELF-IMPROVEMENT IS OURSELVES AS HUMAN BEINGS, AND THE MORE INITIATIVE WE HAVE, THE BETTER."

enrolling in PhD programmes in countries such as Germany, France and Japan — even occasionally at the Hebrew University. Most of the doctoral students have an agreement to return to Al-Quds after their training. Sowwan says his contribution is small, but rewarding. "If I am able to open up an opportunity to an individual, I can feel their happiness."

Neuroscientist Mohammad Herzallah, a postdoc at Rutgers University in Piscataway, New Jersey, works from afar with the Al-Quds' Palestinian Neuroscience Initiative (PNI), which he founded in 2009. The initiative currently includes more than 30 students who aim to become scientists, and they carry out research projects under Herzallah's remote guidance.

One project concerns the biology of depression, which has a prevalence of around 30% in the Palestinian territories, one of the highest rates in the world. The PNI gets financial support from private donors and from the US National Institutes of Health. Together with scientists in Germany, Herzallah is now applying for support from the German research ministry to build a PNI lab that allows Palestinian students to investigate differences in electrical activity in the brains of people with and without depression.

Crowded around a long table, the PNI students update Herzallah on their individual progress in weekly Skype meetings, which take place well before dawn breaks in New Jersey.

The discussions are lively, but Herzallah doesn't tolerate anyone who

nation. At a press conference in July, Saidam said that the problem is "undermining the quality of education and research programmes at our universities".

The Israeli organization COGAT (Coordination of Government Activities in the Territories), which is responsible for visas, did not respond to requests by *Nature* for comments about the specific problems of academics but told *Nature* that there has been no change in its visa policy and that "each individual case is examined on its merits".

Birzeit University says that 8 of its 20 foreign faculty members have failed to get visas in the past 2 years, which Abuhijleh describes as a major problem. Having international scholars keeps the relatively isolated universities connected with the world, he says, "but the visa issue makes it so difficult to get and retain them".

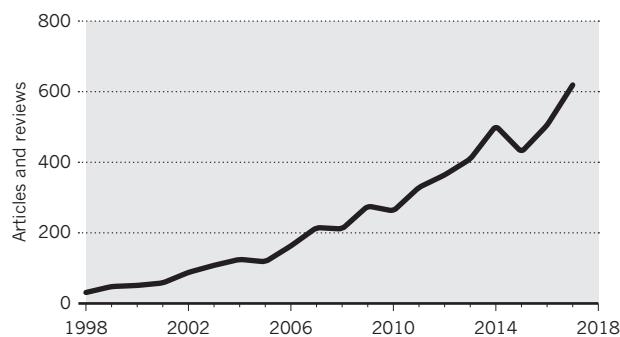
Despite all the challenges, some universities have managed to increase their research output. According to the Scopus database, the rate of scientific publications from Palestinian universities has nearly tripled in the past decade, although the absolute level remains low and many publications are from large international collaborations in the health arena (see 'Science statistics').

Tensions over the occupation and a lack of progress towards peace cast an omnipresent shadow over scientists in Israel, where government money for research and scientific output is above the EU average.

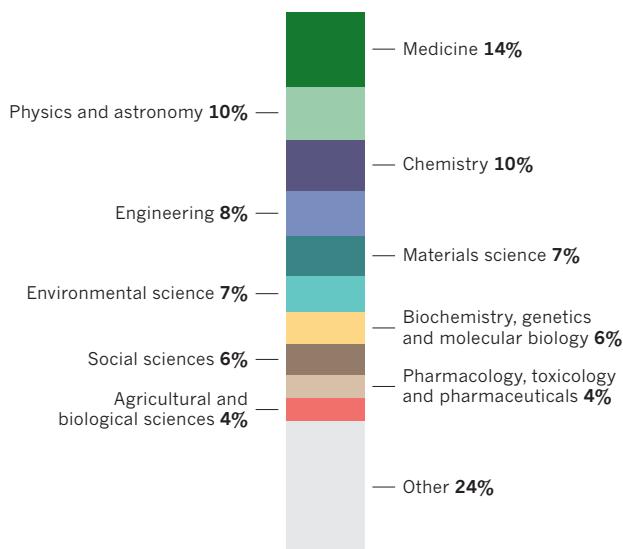
SCIENCE STATISTICS

In the past 20 years, there has been a sharp rise in the number of publications in the Scopus database from scientists in the Palestinian territories, covering a wide range of subject areas.

PUBLICATIONS



SUBJECT AREAS



Israel is second only to Switzerland in the number of prestigious grants its scientists win from the European Research Council, in proportion to its population. Scientific organizations and institutions, including universities, remain carefully neutral on politics, and many individual scientists prefer to keep their heads down, too.

Not all are so reticent, however. Physicist Eli Pollak at the Weizmann Institute of Science in Rehovot, who is also a member of the Israel Council for Higher Education, says that the Palestinian terrorist attacks following the Second Intifada, and the continuing, if lessened, danger, justifies the careful vigilance of Israel's defence system. And the Israeli public has shown strong support for enhancing security measures. Prime Minister Benjamin Netanyahu's party won the highest number of votes in the 2015 election on a platform that emphasized security.

But particularly sensitive government decisions that touch on academia reveal significant splits within scientific organizations. Computer scientist David Harel, also at the Weizmann Institute, is a vocal critic of Israeli politics, which he says is creating an apartheid state that segregates Palestinian and Jewish people and denies Palestinian people many of the most basic rights.

Human-rights activists have heavily criticized the Israeli government's policies in the Palestinian territories, and there have also been concerns about how Palestinian leaders have treated people there.

The organization Human Rights Watch reported last month that the Palestinian Authority in the West Bank and the Islamist militant group Hamas in Gaza crush dissent through violence and imprisonment, which limits free speech in those areas.

In Israel, several government actions in the past year have triggered protests by academics. Education minister Naftali Bennett shepherded through a law in February that would bring a new university in the West Bank's largest settlement town, Ariel, under the umbrella of the Israeli Council for Higher Education. The proposed law was unprecedented and of high political significance, because Ariel is outside Israel's sovereign borders, and some see it as an opening towards annexation of the town.

Harel — who is vice-president of the Israel Academy of Sciences and Humanities — and some other academy members considered the move to also be a danger to the well-being of science in Israel because it offended international colleagues. Harel argued that the academy should issue a dissenting statement, but academy president Nili Cohen disagreed. "It is a political decision of the Israeli government, and the academy does not take positions on political issues," says Cohen, a lawyer at Tel Aviv University.

Harel organized a public letter that was signed by 51 of the academy's 115 members, warning that the controversial law might fuel a tightening of the international boycott of Israel's academia. The legislation was passed anyway, on 12 February.

In another recent conflict, science minister Ofir Akunis blocked the nomination in July of neuroscientist Yael Amitai to the scientific committee of the German-Israeli Foundation for Scientific Research and Development (GIF), which distributes around €12 million per year to collaborative research projects. Akunis says he ruled Amitai out because she had signed a 2002 petition supporting university faculty members and students who refused to carry out military service in the Palestinian territories. The affair caused a storm of protest in Israel and Germany. In Israel, more than 1,300 faculty members signed a petition protesting that the move was political interference in scientific affairs, and calling for a boycott of the GIF until Amitai is appointed. Two German GIF science-committee members resigned.

The Association of University Heads in Israel appealed to the High Court of Justice to decide whether Akunis was within his rights to block her. Akunis has stated that he did so not because of her opinions, but because the petition encouraged "conscientious objection to enlistment in the Israeli Defence Forces". On 11 November, the court ruled that the appeal against Akunis' decision has merit. It will hold a hearing on the case next month, and Akunis cannot appoint a different person until the case is decided.

Occupation politics complicates life for Palestinian scientists — but so does the limited and sporadic nature of financial support. Mutaz Al-Qutob, a chemist who operates a lab in a corner of Al-Quds' elegant central courtyard, says he cannot stick to a clear research agenda, but has to adapt to whatever occasional funding opportunity might appear, to which his ageing equipment might be applied. He struggles to run his mass spectrometer, which was bought with a German grant, and his fish tanks for toxicology studies because he cannot find funds for maintenance and repair costs. He has participated in EU projects on biodiversity in a World Heritage village near Bethlehem, and he has analysed heavy metals in local water supplies, contamination that results in part from the informal e-waste recycling economy that has emerged in the territories. "We are not free to do as we like," he says. "We hope for a better future."

On top of all this, scientists have to cope with the many inconveniences of the region's impoverished environment — not least the poor general infrastructure, including unmaintained roads with densely potholed surfaces that slow traffic even without the border queue-ups.

That is something that Saeed will have to confront less frequently in the new year. She has just acquired a permit that allows her to stay overnight in west Jerusalem, freeing her from a daily commute. One step at a time, she is moving towards her goal "to become a postdoc, to become a professor — if not in Palestine, then anywhere". ■ **SEE EDITORIAL P.293**

Alison Abbott is Nature's senior European correspondent.

A QUESTION OF CONTROL

Clinical-trial participants and their carers are gaining influence over how experiments are run. As they take to social media, that could make things messy for the science.

BY HEIDI LEDFORD

Amber Sapp was browsing the Internet late one night in August when she happened to find out that her 12-year-old son's clinical trial had failed.

Every four weeks for two-and-a-half years, she had shuttled Garrett to a hospital nearly six hours away. There, he was prodded and pricked with needles in the hope that the antibody treatment being tested would reverse a devastating genetic disease called Duchenne muscular dystrophy. But an early data analysis, Sapp learnt, had shown that the treatment wasn't working.

The thought of wasting Garrett's limited time with a failed trial was hard enough. The news was all the more disturbing because it didn't come from the trial organizers, but through a Facebook post from another parent. "It was upsetting that we found out that way," says Amber. "It sent everybody on Facebook into a tizzy." Even Garrett's local clinical-trial coordinator, someone who should have had intimate knowledge of what was happening with the research, hadn't yet heard the news.

Some members of the Facebook group had regularly discussed how their children were faring in the trial, even speculating as to who was in the control arm of the study, receiving a placebo instead of the experimental treatment. Social-media interactions can empower those living with disease, and their families, to make informed choices about their health care and clinical trials. Some people have even united on social media to launch trials of their own.

It's part of a major shift in clinical research. A 2016 survey found that three out of every four major pharmaceutical companies had used a patient-advisory board to gather feedback on clinical-trial designs (S. Stergiopoulos *et al. Ther. Innov. Regul. Sci.*, in the press). And several scientific journals, including *The BMJ*, have included patients as peer reviewers of submitted manuscripts.

But Amber's experience also shows how trial participants are disrupting the usual flow of information in clinical studies. As participants become more empowered, the natural tensions between their goals and those of the researchers become more pronounced. Online discussions threaten to compromise trial integrity when participants join forces to work out who is receiving a placebo. Discussing potential side effects can

also influence results, particularly when the symptoms are subjective. Drug companies have yet to report any cases of such actions causing irrevocable damage to a trial, but some researchers worry that information-sharing by participants could sink trials or weaken their findings.

Now, scientists are grappling with how best to work with — and for — the people they are trying to study. "The fallback for most researchers is, 'I have to get these patients to change,'" says Craig Lipset, head of clinical innovation at Pfizer, a pharmaceutical company based in New York City. "But I think there are other things we'll have to take more seriously in the design of studies."

TRIALS AND TRIBULATIONS

By the time Garrett turned three, Amber, who works as a physical therapist in Nashville, Tennessee, noticed that something was off. When he tried to jump, he couldn't get his feet off the ground, and he looked unstable climbing stairs. Amber asked Garrett's paediatrician for answers, but was told that, in time, he would probably catch up with his peers.

One day, she watched Garrett stand up from sitting on the floor, and the answer came to her. The way that he used his arms to help raise his body was not just a quirk: it was a hallmark of muscular dystrophy that she had studied in school. "It just took me out of the blue," she says. "I thought, 'Oh my God, that's what it is.'"

Duchenne muscular dystrophy (DMD) is a genetic disorder that affects mainly boys, and is caused by mutations in a gene called *DMD*. The dystrophin protein that it encodes is important for maintaining healthy muscle cells; without it, muscles gradually deteriorate. Many people with the disorder need a wheelchair by the time they are 12, and will have difficulty breathing by their late teens.

Amber and her husband spent the next four years consumed by grief. "We refer to them as the dark days," she says. "We couldn't do anything: couldn't function, couldn't talk to other parents, couldn't reach out for resources."

When Garrett was about seven, Amber began to open up. She ventured online and met other carers, chatting to parents of older boys who were grappling with later stages of the disease. "Watching them go through that process of clinical trials and the difficulties — I guess maybe that's where we learned about clinical trials," she says.

Medical centres and pharmaceutical companies have noticed the power of social media to draw in patients. Some have launched efforts to advertise trials, for example to targeted Facebook groups. The hope is that it could help trial recruiters to tackle a growing problem: a shortage of participants that has been stretching the time required to do clinical research.

As companies increasingly focus on rare diseases and precision medicine tailored to a specific subset of patients, it has become more difficult to find willing volunteers who meet the necessary criteria. Recruitment and retention rates are the worst that they've been since the Tufts Center for the Study of Drug Development started tracking them 20 years ago, says Kenneth Getz, who studies clinical trials at the centre in Boston, Massachusetts.

"Industry-wide, everybody recognizes this as a huge problem," says James Nolan, chief executive at InClinica, a contract-research organization in Wayne, Pennsylvania, that conducts clinical trials. "It's not going away — it's going to get much worse."

The recruiting problem has given potential participants leverage and altered their relationship with clinical researchers: a trial that is too burdensome, or forces many participants into a control group, could be doomed to failure from the start. "Many of the companies understand that we can't do this now without patients being equal partners," says Sohini Chowdhury, deputy chief executive of the Michael J. Fox Foundation for Parkinson's Research in New York City.

**"WE COULDN'T DO ANYTHING:
COULDN'T FUNCTION,
COULDN'T TALK TO OTHER
PARENTS, COULDN'T REACH
OUT FOR RESOURCES."**



Amber Sapp looks to many places for guidance on experimental treatments for her son, Garrett, who has Duchenne muscular dystrophy.

So drug firms and medical centres have enlisted the aid of patient advisory boards to evaluate clinical trials. Participants are getting the opportunity to demand trials with fewer procedures, or more comfortable conditions. Lipset recalls a protocol for a trial in atopic dermatitis, a form of eczema, that would have required participants to stop using all their usual medications for six weeks to clear their system of drugs. A panel of people with dermatitis was shocked: going that long without relief was unfathomable. "The washout period made perfect sense scientifically," Lipset says. "But to the humans involved it was completely intolerable."

The team adjusted the protocol, rather than risk launching a trial that was destined to fail. An evaluation of 30 patient advisory boards found that many were making recommendations about the convenience and feasibility of study visits, and the schedule of procedures performed (A. Anderson and K. A. Getz *Ther. Innov. Regul. Sci.* **52**, 469–473; 2018). The advisory boards have good cause to push back. Getz says that as many as one-third of procedures — such as blood tests or biopsies — performed during clinical trials are not crucial to the applications for drug approval.

"Part of the balance is recognizing that although good science is great, it also has to be feasible and convenient," says Getz. "That's where patient engagement has completely changed the philosophy."

In some cases, patients and their advocates band together to launch clinical studies of their own. When Katherine Leon had a heart attack in 2003, soon after the birth of her second child, she was told that it was just something that can happen after having a baby. But Leon eventually learnt that she had spontaneous coronary artery dissection (SCAD), a rare condition that few community physicians are familiar with.

Leon says that she was "randomly googling around" one night when she stumbled on a message board for women with heart disease. Over time, a community of people with SCAD emerged. Then she started keeping a record of their symptoms and disease course: at what age were they diagnosed, which artery was affected and whether it might have been related to pregnancy. She took her data to a physician and convinced her to launch a research project to catalogue

a lot of time with our boys. Nobody has time to waste."

Trial participants have long sought to avoid being in the placebo group; they would rather have the chance to benefit from an experimental drug. The advent of social media has made it much easier to 'unblind' a study, says Pat Furlong, founding president and chief executive of Parent Project Muscular Dystrophy, an advocacy group based in Hackensack, New Jersey. "Before social media, you wouldn't know the other people in that trial," says Furlong, whose two sons had DMD.

Bioethicist Lindsay McNair first became aware of the phenomenon while working for Vertex Pharmaceuticals, which is now in Boston. The company was running a clinical trial of a potential treatment for the hepatitis C virus in 2007 when a researcher reported activity from its participants on MedHelp.org, a health-related social-media site. Some participants said that they were having their blood tested by an outside laboratory to find out their levels of virus, to guess who was receiving the active drug and who the placebo.

McNair, who is now the chief medical officer at WIRB-Copernicus Group in Boston, a company that performs ethical reviews of clinical trials, decided to take a closer look with her colleagues. They read publicly available online health discussions over the course of about a year, noting any that might affect a study's outcome. They found that participants were comparing the appearance and taste of their pills, even crushing them up to get a better look (S. W. Glickman *et al. J. Empir. Res. Hum. Res. Ethics* **7**, 71–80; 2012). Some of the activity, McNair recalls, was on Yahoo Finance company message boards — and at least one financial analyst cited data from these boards in his or her predictions about the trial and in recommendations about Vertex stock.

There is no evidence that online unblinding affected any of these trials. But anecdotes such as these are troubling drug-makers. "We have largely turned a blind eye to the use of social media," says Lipset. "It's only a matter of time before Facebook jeopardizes the scientific integrity of a study."

Sharon Terry, president and chief executive of the advocacy group Genetic Alliance in Washington DC, recalls working on a 2013 trial testing high doses of magnesium in 44 people with a rare genetic disease, pseudoxanthoma elasticum, which affects elastic fibres in connective tissue. "The group of individuals all got on Facebook and figured out pretty fast which were in the control," she says.

In some online conversations that Furlong and McNair have seen, parents discussed leaving a trial if they didn't see any improvement. "Dropouts are super frustrating," says Brian Loew, founder and chief executive of Inspire, a social-media site that caters to people with medical conditions and their carers. This can delay the completion of a trial and raise warning flags to reviewers at regulatory agencies.

And when participants share details about which side effects they might be experiencing, they can induce others to wonder about — and then perhaps report — similar symptoms. The same could be true of a key clinical endpoint of the trial, particularly if that endpoint is somewhat subjective, such as a ranking on a pain scale. And, sometimes, participants swap information about entry criteria, such as the score on a cognitive test that might be required to join an Alzheimer's disease study, says Lipset. Armed with that knowledge, those who want to join the study can prepare accordingly.

Amber says that she generally stayed quiet during such online discussions, but it was still painful to see other families talking about possible improvements in their sons' ability to walk or climb stairs. Garrett had experienced no such progress.

After the first clinical trial, the family began shuttling Garrett to Cincinnati, Ohio, for the antibody trial. The drive, the needles and the time spent in the hospital all took their toll. "Clinical trials are exciting, frustrating and frightening," says Furlong. "There is certainly some altruism. But I can say to you — especially in rare disease, especially when so many people with rare disease are children — what you want, as a caregiver, is benefit."

When Garrett turned 11, Amber held her breath. At that age, he would have to give his own assent to remain in the antibody trial. Garrett agreed, but Amber suspects he bowed to his parents' wishes.

"BEFORE SOCIAL MEDIA, YOU WOULDN'T KNOW THE OTHER PEOPLE IN THAT TRIAL."

features of SCAD. "It was huge, because we felt as patients that we had definitely initiated it," Leon says. "When I compare what they've discovered so far with the anecdotal data in my little proposal, it jibes pretty well — and that's all just from people having conversations."

A PLACEBO EFFECT

Garrett's first clinical trial was designed to test whether a drug called tadalafil (Cialis) would help to keep boys with DMD walking. The protocol was relatively simple: just a few pills in the morning with a spoonful of apple sauce.

But Garrett's ability to walk continued to decline. Faced with a degenerative disease and a ticking clock, the family wrestled with worries that he should move on to another trial. Eventually, Amber called the coordinator and said it was time to consider leaving the trial and to look ahead to the next one.

Online, Amber could see carers facing the same decision in various clinical studies. Some parents posted videos of their children walking or climbing stairs, and speculated as to whether they were receiving the active drug. If they suspected that their child was taking the placebo, a number of parents openly talked about their plans to withdraw from a study. "Nobody wants to be in the control," says Amber. "We don't have



Garrett Sapp (right), a 12-year-old who has Duchenne muscular dystrophy, might be eligible for gene-therapy trials. But the decision to enrol is a complex one.

ABIGAIL BOBO

Furlong recognizes that anxiety. “There’s a moment when your son looks at you and says, ‘I don’t think I want to do this. I miss my friends. I don’t want them to stick me another time.’” she says. “As a parent, you are second-guessing: ‘Is this the right thing?’” Often, parents of children with DMD will share information online because they are desperate to hear someone, anyone, tell them that their child is improving, she says.

Researchers are still grappling with how best to handle such online discussions. Inspire, which displays targeted advertisements for clinical trials to some of its 1.5 million members, expressly prohibits discussions that could affect clinical-trial results, such as comparing possible side effects or discussing ways to game eligibility criteria to gain entry to a trial. The site employs moderators to check posts after they go live.

“We had a lot of internal debate about it,” says Loew of the policy. “On the one hand, people should be able to talk about whatever they want. But we decided that you can actually do harm to the science.” Other sites, however, such as Twitter and Facebook, have no such policies.

Some companies running trials have inserted guidance about such communications in the consent forms that study participants sign. But that can backfire and cause undue worry, or limit participants’ ability to find support online, says Lipset. “You can see in online communities where participants are scared that they have just signed a confidentiality agreement and will be thrown in jail for posting.”

Lipset says that investigators will have to become savvier about how they set up their trials. This could include firming up eligibility criteria for a study, he says, to make them less subjective — and harder for a potential participant to game.

Some firms are hiring outside companies that specialize in listening in on social media, to report back when conversations veer towards unblinding a trial. Others are looking to facilitate the groups. Bristol-Myers Squibb, headquartered in New York City, partnered with Inspire to launch a moderated online community in April, in which patients in a given trial can support one another and discuss their condition, says

Loew. This idea is catching on, says Lipset. “We’re maturing to a place where people have to take seriously even the potential to create online communities for your research participants, so that people can have a safe place to share. Because they want to share.”

THE TOUGH DECISIONS

When Amber learnt that Garrett’s second trial had ended, it was time to weigh options for the next one. But Garrett’s choices are narrowing. He stopped walking this summer, and few trials will take boys who are no longer able to walk.

The family then considered a gene-therapy trial. It was a difficult decision. “Gene therapy is huge and promising and terrifying at the same time,” Amber says.

It comes with a slew of new challenges, and risks. The virus that is used to deliver genes could raise an immune response that would make Garrett ineligible for future gene-therapy trials. And if he’s in the placebo arm, he won’t know whether he’s eligible to receive the actual treatment until a year has passed. Added to these tensions would be three muscle biopsies performed under general anaesthesia, procedures that are particularly unnerving for people whose muscle is wasting away. “If the trial we had just come out of was, to us, pretty invasive, this is ten times that discomfort,” Amber says.

It’s a gamble. In October, Amber and her family opted to hold off from joining the gene-therapy trial. While they were weighing their options, Amber decided not to rely on other parents on social media to help with the decision. Instead, she stuck to her “board of directors”, a few trusted medical professionals. “Social media has such a wide pool of people that you don’t always know that the answers you’re going to get are on the level,” she says. “It’s hard,” Amber adds. “Time is limited.” ■ **SEE EDITORIAL P.293**

Heidi Ledford is a senior reporter for *Nature* in London.

COMMENT

HISTORY Samuel Goudsmit, spin physicist, atomic sleuth and journal editor p.320

ARCHAEOLOGY Virtual reality rebuilds architectural rubble, bit by bit p.321

SUSTAINABILITY Governments should unite to cut meat consumption p.325

OBITUARY Kuen Charles Kao, pioneer of optical fibres, remembered p.326

ILLUSTRATION BY DAVID PARKINS



Be open about drug failures to speed up research

Access to evidence from disappointing drug-development programmes advances the whole scientific process, explain **Enrica Alteri** and **Lorenzo Guizzaro**.

Self-help and business books are replete with advice for learning from failures. The biomedical community must do just that if it is to ease the burden from intractable conditions such as Alzheimer's disease. It can take 20 years or more to get a drug

to market, from testing compounds in animals to running late-stage (phase III) clinical trials in thousands of subjects. More than 80% of drugs that are tested in humans fail to demonstrate safety and efficacy¹ (see 'High failure rate'); the rate for Alzheimer's

treatments is estimated at more than 99% (ref. 2; see 'Alzheimer's drug attrition').

Yet the data behind these failures are generally not seen by regulators, or considered deeply by anyone outside the company sponsoring the trial. Without this ▶

► information, learning is unlikely.

In 2015, drug companies were invited to discuss confidential information about all their Alzheimer's disease programmes by the European Medicines Agency (EMA), where we work. An important result of this data-sharing initiative was new recommendations for designing clinical trials and assessing patients' outcomes, as consolidated in EMA's revised guideline for clinical investigations of Alzheimer's disease treatments³. We believe that what the companies learnt (indirectly) from one another will lead to faster, more-informative clinical trials. In our view, if this information had been put together sooner, decision-making after early-stage trials could have been improved.

SHARING SURGE

Practices to enable more-thorough, earlier analyses of failed developments should be adapted to treatments for other challenging diseases, and should be part of regulators' responsibilities. This will ensure that clinical research evaluates treatments faster and with more certainty.

Initiatives for private companies to share biomedical data and ideas have expanded in the past decade. Some, such as the Biomarkers Consortium and the Structural Genomics Consortium, bring together many companies and academics to design experiments for the benefit of the community, such as identifying disease markers or characterizing tool compounds to understand how target proteins work. Others ask companies and academic groups to pool data in a common repository. For instance, the Project Data Sphere Initiative is a platform to share de-identified data from people who were enrolled in the control,

placebo or even experimental arms of more than 180 cancer trials.

More data are also being put into the public domain from individual trials. The International Committee of Medical Journal Editors has advocated for the release of large quantities of data from trials that have had results submitted for publication⁴. For its part, EMA has started publishing all clinical-study reports for medicines after regulatory review is completed, together with its assessment of the preclinical and clinical evidence⁵. Although these data are useful, they do not encompass information for drug candidates that fail to make it to regulatory submission.

At best, some research leading to negative results in clinical trials will appear on clinical-trial registries or, perhaps, in publications, but without the context of how these compounds performed in preclinical or early-stage programmes. Moreover, the time lag between the generation of data and any eventual accessibility is usually very long, hampering efforts to learn.

THE WHOLE STORY

Information that is not shared is arguably the most important: data that failed to meet drug developers' hopes are most likely to help progress. Large clinical trials are multimillion-dollar experiments to validate a hypothesis that an experimental drug will be effective and safe. Results that go against these expectations must be made available to refine hypotheses and to elaborate alternative ones.

Data from negative research can reveal whether a trial adequately tested the intended hypothesis. For example, in cardiovascular disease, three clinical trials of inhibitors of cholesteryl ester transfer protein (CETP)

showed no effect and led to questions over whether CETP was an appropriate target. When a fourth trial of a CETP inhibitor found that it modestly reduced the risk of a coronary event, such as a heart attack or unstable angina, the result led to speculations that the target was indeed promising. The problem arose because of the way in which molecules were tested, and because it was difficult to find molecules that inhibited CETP enough to make a measurable difference. (The company running the fourth trial elected not to pursue that product further.) We have this insight because the CETP cardiovascular trials were all large and disclosed⁶.

Going back to the bench to elaborate a new hypothesis for treating a disease is likely to delay drug discovery by a decade or more, so it is crucial to assess whether there are other ways forward. We in the scientific community wanted to know what could be learnt from earlier, undisclosed work on Alzheimer's.

Alzheimer's disease is perhaps the therapeutic area best positioned to encourage this level of cooperation. In the past 10 years, more than 30 drugs have entered phase III clinical trials for Alzheimer's disease. So far, none of these experimental treatments have shown therapeutic benefit or even met trial objectives, such as halting or reversing the decline in a person's cognition or ability to perform everyday activities. There is evidence that, in some cases, these trials were not preceded by adequate exploratory research. This led to a high rate of failures, increased the risk of researchers missing therapeutic potential even if it existed (for example, by selecting a wrong dose or inappropriate target population), and created a near-certainty of obtaining results that are difficult to interpret.

Many development programmes for Alzheimer's treatments have announced disappointing results: starting in 2012, large, highly anticipated trials sponsored by Merck, Pfizer, Johnson & Johnson, Eli Lilly and Roche all failed to show therapeutic benefits.

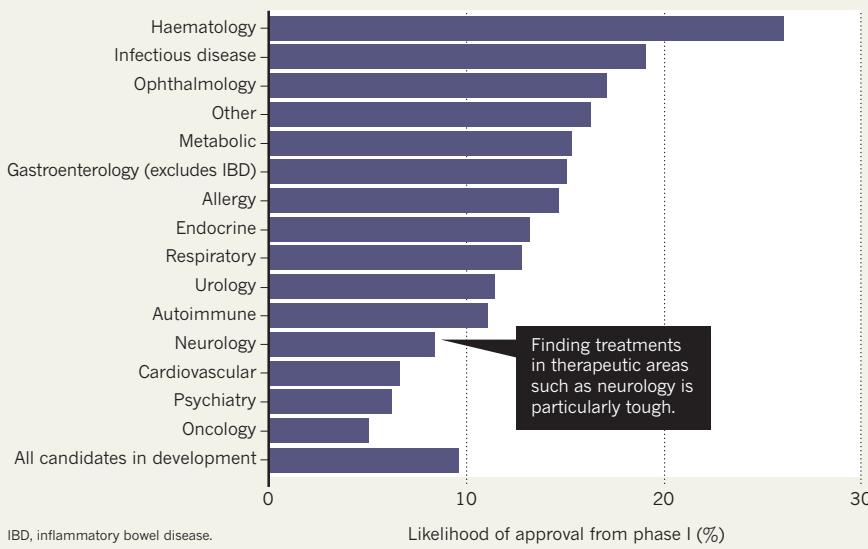
Health ministers from countries in the Group of 8 (G8) published a policy paper on dementia in 2013 that was intended to stimulate action from all players (see go.nature.com/2sg9th2). In 2015, the World Health Organization included "increasing collective efforts in dementia research and fostering collaboration" in its global call for action on dementia. These initiatives, together with the previous failures, meant that drug companies faced significant public pressure to demonstrate that they had taken action towards solutions. Working voluntarily with regulators offered a good way to do so.

PRIVATE POOL

Following the G8 call to action, we at EMA invited drug companies to present their research to us confidentially and individually — detailing what drug targets

HIGH FAILURE RATE

In 7,455 drug-development programmes from 2006 to 2015, fewer than 10% of experimental drugs were found to be safe and effective, and then approved for market.



SOURCE: REF²
they investigated, what populations they thought their interventions might treat, and how they intended to test this in their trial designs. Seven companies agreed to take part. Their presentations to us covered data on 14 discontinued or ongoing trials, including efficacy trials that collectively covered more than 12,000 participants.

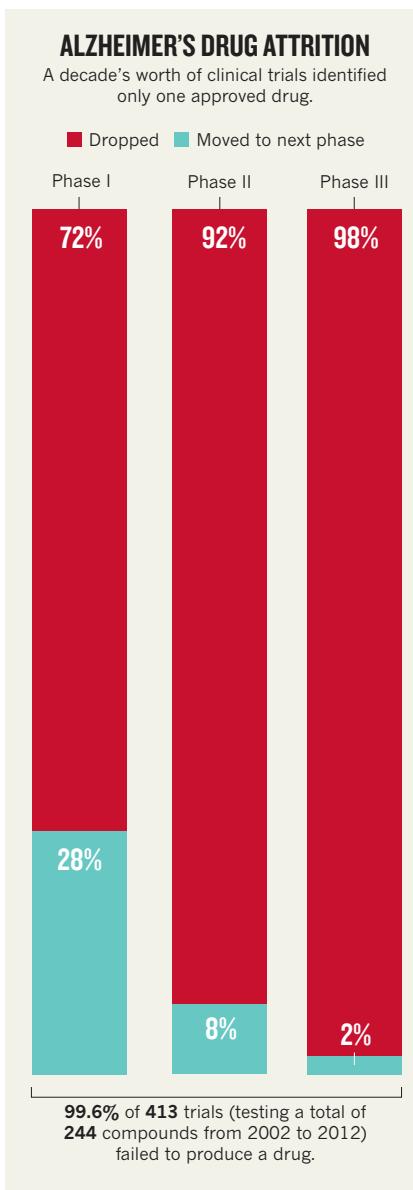
We did not ask them to give us their files so that we could do our own analyses. Instead, we invited them to walk us through their logic and the evidence that led, in most cases, to disappointing results in large clinical trials. The point was not to combine data to perform more powerful statistical analyses, but to review the entirety of data that each company provided and then to consider common issues.

We looked at the landscape of research and development (R&D) plans with the knowledge that pivotal clinical trials had been negative. We considered the hypotheses that the companies put forward originally, how they set up studies to test their hypotheses, how they developed *in vitro* assays and animal models, and how they interpreted signals in early clinical work. We considered what we could learn from one company's studies in light of another's. We tried to understand when ideas about potential therapies went wrong.

The information shared with our teams was more up-to-date, broader and more in-depth than what is commonly published in the literature, included in trial registries or given in mandated public summaries. Details on data generated before phase III trials — including preclinical and early clinical research — were crucial to frame the failure's significance in terms of which hypotheses it falsified. Such information also helped to avoid unwarranted negative conclusions about uninformative generic terms such as 'β-amyloid hypothesis' (the theory that accumulation of the peptide amyloid-β in the brain is what causes the disease), which could encompass multiple molecular targets and strategies.

These insights improved our understanding of the disease, how it progresses and, importantly, how modulating the supposed mechanism of disease might have a clinically detectable effect (unpublished results). For example, how strongly must a potential drug molecule bind to a target protein to alter physiology? What fraction of the molecule must penetrate the blood-brain barrier to have an effect? These parameters become clearer with data from multiple, diverse programmes.

Although we are legally limited as to which data we can present publicly, our work helped us to revise the EMA guideline³ that we think will aid the design of more-informative Alzheimer's trials and better R&D programmes. For example, people enrolled in trials should be assessed both for their symptoms and for



evidence of amyloid-β pathology. This makes the progression of disease more predictable and enhances the power of the trial. This exercise also allowed us to develop recommendations for how to consider 'intercurrent events', such as a stroke or change in medication regimen, that some older participants in a long trial will inevitably experience, and which complicate the interpretation of results.

We also uncovered problems with outcome measurements. Some of the previously used and best-known instruments have proved inadequate to study Alzheimer's disease in its early stages^{7,8}. To overcome this issue, most trials combine single items from various outcome measures, gauging specific aspects of cognitive performance and function in daily activities. Although this strategy has merit, the interpretability of results should take precedence over a purely statistical approach. In addition, different practices across trials limit efforts to compare results. We think that

more-standardized approaches to measuring outcomes in our revised guideline will lead to more-informative trials.

DATA FEED GOOD SCIENCE

How were we able to do this? Because regulators routinely work with commercially confidential information, companies can be willing to share data with regulators that they would be reluctant to put in the public domain.

For this project, EMA worked with the regulatory agencies of Canada, Japan and the United States to align requirements as much as possible⁹. Pharmaceutical companies often claim that different regulatory requirements impede global development. Although this is still a challenge, our focused multilateral effort identified areas for convergence — such as selecting the study population and assessing patient outcomes — that can aid clinical investigations. Such convergence is reflected in the US Food and Drug Administration's revised industry guidance on Alzheimer's disease (see go.nature.com/2jbvsas), which was published shortly before EMA's latest guideline³ and contains similar recommendations.

It is too early to assess whether Alzheimer's drug-development programmes led by the new EMA guideline will yield positive results. Nonetheless, we think our efforts demonstrate that the gains for overall progress in pharmaceutical research should outweigh any individual company's hesitation to disclose data. Furthermore, it shows that regulators can act as enablers of more effective R&D. To speed up progress, companies must be more forthcoming with their data and thinking, and regulators must find ways to help them with this. The ultimate goal is to allow broader access to data from drug-development programmes and to enable faster learning by the entire research community.

We hope that this project leads to similar efforts in other diseases that are difficult to treat. We owe it to the public and to patients to ensure that R&D efforts continue to move towards greater transparency. ■

Enrica Alteri is head of the R&D support division at the European Medicines Agency in London. **Lorenzo Guizzaro** is a scientific officer at the European Medicines Agency. e-mail: enrica.alteri@ema.europa.eu

1. DiMasi, J. A., Feldman L., Seckler, A. & Wilson, A. *Clin. Pharmacol. Ther.* **87**, 272–277 (2010).
2. Cummings, J. L., Morstorf, T. & Zhong, K. *Alzheimer's Res. Ther.* **6**, 37 (2014).
3. European Medicines Agency. *Guideline on the Clinical Investigation of Medicines for the Treatment of Alzheimer's Disease* (EMA, 2018).
4. Taichman, D. B. et al. *Ann. Intern. Med.* **167**, 63–65 (2017).
5. Bonini, S., Eichler, H. G., Wathion, N. & Rasi, G. *N. Engl. J. Med.* **371**, 2452–2455 (2014).
6. Tall, A. R. & Rader, D. J. *Circ. Res.* **122**, 106–112 (2018).
7. Benge, J. F., Balsis, S., Geraci, L., Massman, P. J. & Doody, R. S. *Geriatr. Cogn. Disord.* **28**, 63–69 (2009).
8. Jekel, K. et al. *Alzheimer's Res. Ther.* **7**, 17 (2015).
9. Molzon, J. A. et al. *Clin. Pharmacol. Ther.* **89**, 503–512 (2011).



Samuel Goudsmit at a dinner at the University of Michigan in Ann Arbor in the 1930s.

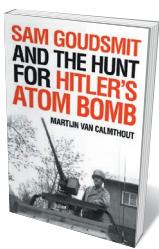
HISTORY

From quantum spin to wartime spy

Davide Castelvecchi enjoys a life of Samuel Goudsmit, the atomic sleuth nominated for a Nobel prize 48 times.

In 1945, four months after the end of the Second World War in Europe, a 40-something man drove a US Army Jeep through the ravaged streets of The Hague in the Netherlands, where he had grown up. When Samuel Goudsmit stepped into his parents' house, he found it partly dismantled — probably scavenged for wood. He knew he would not find his parents here; two years before, he had received a farewell letter from his mother Marianne, postmarked Westerbork. This was the transit camp where she and many other Dutch people of Jewish ancestry, including Anne Frank, were held on the way to Nazi extermination camps such as Auschwitz.

This scene is a fitting start to the gripping *Sam Goudsmit and the Hunt for Hitler's Atom Bomb* (first published in Dutch in 2016, now translated into English by Michiel Horn). Author Martijn van Calmthout (former science editor of Amsterdam-based newspaper *De Volkskrant*) argues that the wartime years



Sam Goudsmit and the Hunt for Hitler's Atom Bomb
MARTIJN VAN CALMTHOUT
Prometheus (2018)

along with the painful knowledge that he wasn't in the same league as other quantum pioneers.

Goudsmit stumbled to fame in 1925. For more than a decade, Niels Bohr and others had been trying to develop a quantum theory

are the key to understanding Goudsmit's extraordinary life. His achievements spanned the co-discovery of quantum spin, the successful search for Adolf Hitler's atomic scientists and the founding of pre-eminent physics journal *Physical Review Letters* — as well as at least 48 nominations for a Nobel prize. Goudsmit overcame significant hardships,

of the atom, mostly by studying atomic spectra. These consist of the energies (specific to each element) of the quanta of light, or photons, that an atom's electrons can absorb or emit. Physicists had been struggling to make sense of anomalies that appeared in spectra when atoms were immersed in a magnetic field: some spectral levels mysteriously split into two or more. Goudsmit and his friend George Uhlenbeck, both graduate students at Leiden University in the Netherlands, had an idea.

They proposed that the splitting could be explained if the electron had an intrinsic 'spin' that could assume one of two directions: clockwise or anticlockwise. Other physicists had discarded this idea, seeing it as marred by conceptual difficulties. For instance, it seemed to imply that electrons should rotate faster than the speed of light. Goudsmit and Uhlenbeck were blissfully unaware of all that. (And more: when Uhlenbeck suggested spin as an extra degree of freedom for the electron, Goudsmit asked, "What is a degree of freedom?") Their paper was published that year (G. E. Uhlenbeck and S. Goudsmit *Naturwissenschaften* 13, 953–954; 1925).

Soon, researchers including Paul Dirac explained away the conceptual difficulties. Quantum spin was born. It is one of the basic properties of all subatomic particles, and a crucial step to understanding the periodic table: without it, atomic structure would be completely different. Yet Goudsmit and Uhlenbeck never received a Nobel prize for their discovery, perhaps because the idea had already been discussed by others, including physicist Ralph Kronig.

After graduation, the two researchers emigrated to the United States, setting the stage for Goudsmit's participation in the US war effort. German scientists discovered nuclear fission in 1938. During the war, Allied forces feared that Hitler might be close to having an atomic bomb. So, in 1944, as D-Day approached, Leslie Groves — the general in charge of the Manhattan Project, the United States' own effort to build the bomb — organized an intelligence mission to follow the Allied invasion. It would scour Germany for clues to the country's nuclear efforts, and apprehend its leading nuclear physicists.

Goudsmit was picked as scientific leader. As van Calmthout argues, Goudsmit was perhaps uniquely qualified for the job. Physics was a small world then, and he knew most of the potential suspects personally, notably his friend Werner Heisenberg, founder of quantum mechanics and undisputed leader of Germany's physics community. Goudsmit also had an investigator's nose: he grew up reading detective stories, and had even considered a career in forensic science.

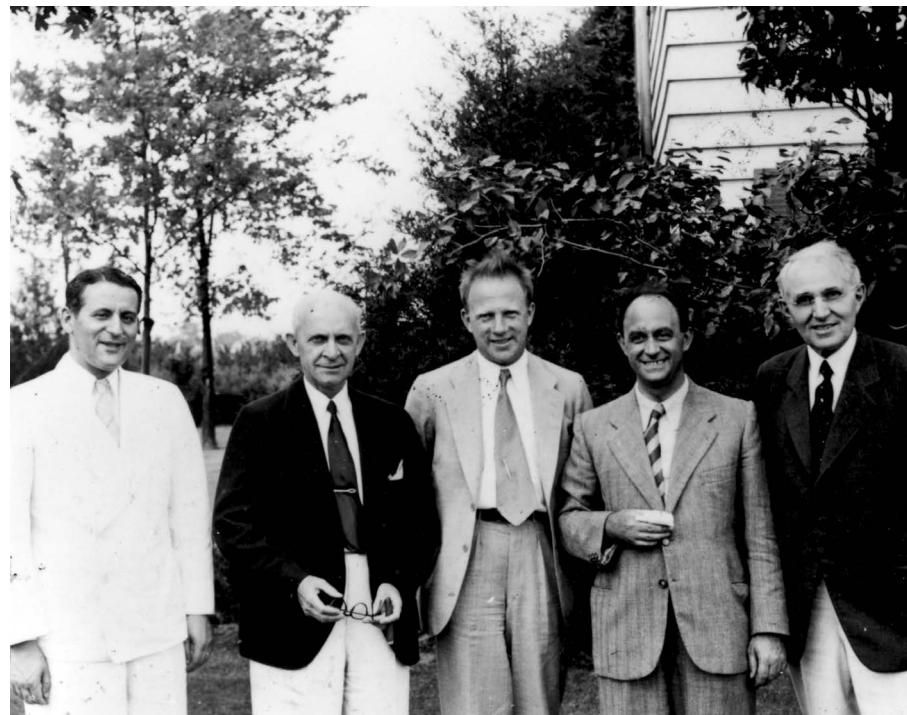
If van Calmthout's account of the mission reads like a thriller, that's because it was. It culminates with the discovery of

Heisenberg's lab in Haigerloch, Germany, where, in hiding, he had tried and failed to get a primitive nuclear reactor started in a former beer cellar. When US soldiers walked into Heisenberg's office, they found a photo of him taken in Michigan in 1939. Goudsmit was in it, too: he ran the summer school that Heisenberg was visiting at the time.

Goudsmit told this story in his spell-binding 1947 memoir, *Alsos* (the mission's code name). But van Calmthout's narrative is hugely enriched by details from other sources. These include letters from Goudsmit to his first wife, Jaantje, and now-declassified documents. Key among these are the transcripts of recordings collected by British intelligence while eavesdropping on Heisenberg and fellow physicists during their internment in a Cambridgeshire country house, Farm Hall (see A. Finkbeiner *Nature* **503**, 466–467; 2013).

Goudsmit realized as early as November 1944 that the Nazis' nuclear 'programme' never amounted to much. The question of why not is still controversial, and van Calmthout does a good job of describing its subtleties. One thing is clear. The 'official' version that Heisenberg presented postwar — that they could have built a bomb, but decided not to — became untenable after the Farm Hall transcripts were declassified in the 1990s. Those show that some of the interned scientists even mocked Heisenberg for being a "second-rater".

Although van Calmthout has a background in physics, *Sam Goudsmit* is not a scientific biography. It devotes little space to the intellectual development of ideas during what was the most momentous period in



L-R: Samuel Goudsmit, Clarence Yoakum, Werner Heisenberg, Enrico Fermi and Edward Kraus in 1939.

physics history so far. This makes the book accessible. But it has few references and no notes. Van Calmthout taps his source material liberally, but is coy on their details.

Goudsmit's later years might seem anticlimactic. He died in 1978; before that, he lived comfortably as a high-level official in a US national lab and as editor-in-chief of *Physical Review* and its spin-off, *Physical Review Letters*, which he founded in 1958. All along, he complained about how Big Physics had

changed the field by necessitating expensive machinery. But his later achievements perhaps hold a lesson for an era in which despotism is once again on the rise globally. They show how Goudsmit's generation of scientists managed, despite the depredations and cruelties of Nazism, to persist long after the Reich had fallen. ■

Davide Castelvecchi is a physical-sciences reporter at *Nature*.

ARCHAEOLOGY

Ancient cities rescued from rubble, bit by bit

Laura Spinney turns virtual tourist among digital reconstructions of monuments destroyed by war.

For more than 800 years, a minaret dominated the skyline of Mosul, Iraq. Nicknamed *al-Hadba*, or 'the hunchback', because of its 3-metre tilt, it belonged to the Great Mosque of al-Nuri, commissioned in the twelfth century. Mosque and minaret were reduced to rubble after Islamist terrorist group ISIS took the city in 2014.

Today, both can be seen in an exhibition at the Arab World Institute (AWI) in Paris. The reconstruction is digital, not physical,

Age Old Cities
Arab World Institute,
Paris. Until 10
February 2019.

but the translation of the former into the latter is under way in now-liberated Mosul, thanks to a 5-year, US\$50-million rebuilding project announced this year. The exhibition aims to show how digital technologies are redefining rescue archaeology and contributing to the preservation of our past.

The Monumental Arch of Palmyra in Syria, destroyed by ISIS in 2015, was recreated first digitally and then in Egyptian

marble — in which form it is currently touring the globe. That initiative was criticized for stripping the arch of its context. This exhibition avoids that error, and gives only a nod to the arch, the best-known product of digital archaeology so far.

The show focuses on four sites of historical importance in the Arab world: Mosul and Palmyra, along with Aleppo in Syria and Leptis Magna in Libya. All have seen empires rise and fall; all sit atop layers of rich archaeological material. Some are also, as this exhibition reminds us, living cities.

On a giant screen in the first room, the Old City of Mosul is projected in three dimensions. A fly-over view shows how ancient monuments are embedded in urban fabric, and how badly both have been damaged. Before our eyes, the monuments are rebuilt, virtually. Paris-based start-up Iconem partnered with the AWI to create these dense projections by combining data from aerial images taken by drones, pictures taken at ground level using a boom, and old photographs of the monuments before ►

► they were destroyed. The drones enabled Iconem's team to penetrate cities before they had been de-mined.

There are no physical objects in the exhibition. Through video documentaries and interviews, visitors learn that until very recently, several religious groups lived side by side in Mosul. One of the symbols of the city's multiculturalism was the tomb of Jonah — a prophet for Jews, Christians and Muslims — who was buried in the ancient Assyrian city of Nineveh, where Mosul now stands. After the city was liberated in 2017, Iraqi archaeologists discovered that ISIS had destroyed Jonah's tomb and dug a network of tunnels beneath. Exploring these, they stumbled on the remains of an Assyrian palace. Meanwhile, the previously hidden remains of a synagogue were discovered in the rubble of the Jewish Quarter.

Aleppo's story is different. The damage there was collateral, the fallout from fighting between the Syrian regime and rebel forces between 2012 and 2016. At the centre of old Aleppo lies a magnificent thirteenth-century citadel, itself built over remains from the Roman period or even earlier. Syrian troops made the citadel their base; from here they bombarded the rebels in the city beyond, so it came through relatively unscathed. The city's souks, on the other hand, were badly damaged. Commerce is Aleppo's beating heart, and the market-places are being rebuilt. Exhibition visitors can wander virtually through them as they once were.

The 3D reconstructions of Palmyra and Leptis Magna face each other across a room, because of what they have in common and what sets them apart. Both are purely archaeological sites, not embedded

The decision to destroy and the decision to rebuild are political.



In a 3D reconstruction, a damaged souk in Aleppo, Syria, rises from its own rubble.

in modern cities; but ISIS destroyed 80% of Palmyra, although it mostly spared the site's Roman theatre, where the group staged its executions. The lesser-known Roman site of Leptis Magna — dubbed the Rome of Africa — has been looted and neglected, and is threatened by rising seas. The pairing makes a larger point: war isn't the only threat to our material heritage, nor the only one to which digital reconstructions provide at least a partial response. Donning a headset, visitors can become virtual-reality tourists and see that for themselves.

There is one glaring omission that makes

this poignant exhibition even more timely. No mention is made of the Yemeni sites damaged by Saudi bombs, such as the almost-4,000-year-old Marib Dam. The AWI is partly funded by Saudi Arabia. And the omission underlines the fact that both the decision to destroy and the decision to rebuild are political — as Warsaw, Coventry and Dresden, ravaged in the Second World War, know only too well. ■

Laura Spinney is a writer and science journalist based in Paris.
e-mail: lspinney@gmail.com

ENVIRONMENTAL RE-ENGINEERING

Lake Lazarus: rewilding the US west

Amy Maxmen lauds a study on a bold project to re-engineer a dry lake bed.

At the start of the twentieth century, Owens Lake in southern California was one of the largest inland bodies of water in the United States. By the mid-1920s, it was gone, drained to provide water to a mushrooming Los Angeles. Over the past 30 years, the city has spent around US\$2 billion to undo the damage. It has failed to restore the lake, but in *The Spoils*

The Spoils of Dust: Reinventing the Lake that Made Los Angeles

ALEXANDER ROBINSON
Applied Research & Design (2018)

of Dust, Alexander Robinson describes how the effort has succeeded in another way: by creating a landscape no less valuable ecologically. By documenting the transitions

the lake has undergone, he suggests a way forward for engineers, geologists, ecologists and landscape designers hoping to bring other environments back from the brink.

The despoiling of the lake (which was nearly the length of Manhattan, New York) began in 1913. Former president Theodore Roosevelt had ratified a plan for an aqueduct that would divert water from the

Sierra Nevada mountains to Los Angeles, instead of the lake. Within two decades, the city's population had more than quadrupled. By then, the lake bed was dry and the city sought supplies elsewhere. Winds cascading off the mountains swept up storms of dust from the barren land. Sulfate salts eroded clay soils, and toxic particulate matter, including arsenic and cadmium, wafted into the atmosphere. Scientific studies concluded that the dry lake bed was causing itchy throats, burning eyes, asthma and other respiratory problems in the surrounding communities.

In 1990, this human-made dust bowl prompted the US Congress to amend the 1963 Clean Air Act to include land use, as well as industries, as sources of pollution. At times, the measure of particulate matter at Owens Lake was the highest in the country, at more than 120 times the Environmental Protection Agency's air-quality limit. The Los Angeles Department of Water and Power decided that the most expedient solution was to refill a large portion of the lake.

Unsurprisingly, the plan failed. The land has been altered irrevocably, and only shallow, temporary pools formed along the basin. However, the dust storms did die down, and the city avoided regulatory fines. So it continued to dampen the dust by flooding the area and adding gravel.

Initially, the authorities paid little attention to restoring local ecology, but the transformation of the dry lake bed into a heterogeneous expanse dotted with saline pools encouraged the return of wildlife. In 2010, nearly 40,000 native and

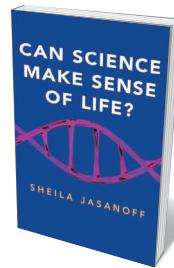
migrating birds were counted in a single day, including rare western snowy plovers (*Charadrius nivosus nivosus*).

In 2011, the number was closer to 60,000; by 2013, it was up to 115,000. And nearly 5% of the world's population of American avocets (*Recurvirostra americana*) were seen at the lake in 2013.

The same year, the city proposed putting up to \$1 billion into the dust-control project, which now included habitat, cultural resources and economic development among its goals. The new iteration of the project prompted an unprecedented level of data collection. NASA satellites that measure short-wave infrared bands were calibrated with tap tests on the ground to track wetness across the expanse. Geographers used the Global Positioning System to map topographical features. And bird-watchers — professional and amateur — flocked to the site. Intermittent flooding created microhabitats for an enormous

"Intermittent flooding created microhabitats for an enormous diversity of birds."

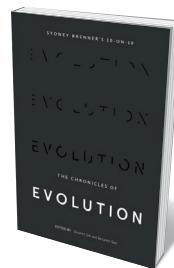
Books in brief



Can Science Make Sense of Life?

Sheila Jasanoff POLITY (2018)

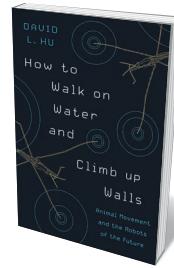
From gene drives to synthetic organoids, every rapid advance in the life sciences opens up a hot-button issue. This incisive study by sociologist of science Sheila Jasanoff examines ethics at that cutting edge. She argues that the view of the human genome as a 'book of life', read primarily by biologists, is partial; alongside it belong fields such as ecology, which explore what life is, rather than what it is for. Interweaving cultural touchstones, science history and trenchant insight, Jasanoff calls for a biology that reintegrates humanistic concerns to prevent a reductionist scientific hegemony.



Sydney Brenner's 10-on-10: The Chronicles of Evolution

Edited by Shuzhen Sim and Benjamin Seet WILDPY (2018)

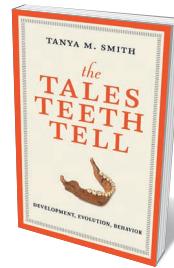
Spanning 14 billion years and 10 timescales, this scientific chronicle (brainchild of Nobel-prizewinning geneticist Sydney Brenner) addresses the monumental question of how humanity has come to dominate Earth. Among the 24 prominent scientists and thinkers who contribute are mathematician John Barrow on the habitable zone, biotechnologists Giulia Rancati and Norman Pavelka on cellular complexity, neuroscientist Atsushi Iriki on the evolution of human higher cognition and social scientist Helga Nowotny on our "radically open future". A lavishly illustrated, thought-provoking ride.



How to Walk on Water and Climb Up Walls

David L. Hu PRINCETON UNIVERSITY PRESS (2018)

Animals are the ultimate movers and shakers, proves biomechanical engineer David Hu in this engrossing tour of faunal motion. Hu reveals propulsive genius in myriad beasts: a mako shark's 6-metre leap from the sea; a skink that swims in sand; the sinuous 'flight' of a *Chrysopelea* gliding snake. Even the *Periplaneta americana* cockroach does more than scuttle at lightning speed: its structural similarity to a stress ball allows it to withstand severe pressures. And the physical principles unveiled, Hu shows, offer as much to fluid dynamics and robotics as they do to evolution and zoology.



The Tales Teeth Tell

Tanya M. Smith MIT PRESS (2018)

Biological anthropologist Tanya Smith drills into what disinterred teeth, as "sophisticated time machines", can tell us about individuals, our species and the deep past. Her study — technically chewy yet thoroughly engaging — examines the human story through dental development, evolution and related behaviour, interlacing vivid anecdotes from her scientific career. The result is a mix of fascinating findings at all scales, from scanning electron microscopy displaying the exquisite geometry of enamel prisms, to toothpick use among hominins some 2 million years ago.



The Continent of Antarctica

Julian Dowdeswell and Michael Hambrey PAPADAKIS (2018)

Part-paean, part-study, this many-faceted portrait of Antarctica meshes crisp scientific writing with luminous images. Julian Dowdeswell — director of the Scott Polar Research Institute in Cambridge, UK — and glaciologist Michael Hambrey examine the continent through lenses from the geographical to the biological, touching, too, on its role as home to a shifting population of researchers. Drawn from years of fieldwork, this is a book sparking renewed awe over this stupendous landmass, outpost of the climate system and — with the sea bed — Earth's final frontier. **Barbara Kiser**



WITOLD SKRYPCZAK/ALAMY

California's Owens Lake, once one of the largest inland bodies of water in the United States, shrank to nearly nothing in the early twentieth century.

► diversity of birds: shorebirds, divers and migratory fowl were drawn to various salinities, depths and the particular vegetation or invertebrates in each niche.

Throughout *The Spoils of Dust*, Robinson, a landscape architect at the University of Southern California in Los Angeles, points out that lakes need not be appreciated only as static bodies of water — a dictionary definition. Owens Lake no longer fits that definition; but nor did it 4,000 years ago, when the area was an arid expanse. By revealing the return of life to the site, Robinson invites the reader to appreciate this landscape as a phase in the lake's progression, rather than see it as a wasteland. After all, to embrace this new environment, the public has to like it.

The Los Angeles city government has come around to this view. In the past few years, it has included artists and landscape architects in the restorations of Owens Lake. For example, an open plaza on one side of the expanse is meant to resemble birds' wings; etchings in a dust-control zone read "Tweet, Tweet" when viewed from above. But Robinson says that this municipal effort, although moving in the right direction,

misses the mark. Instead, designers might work with features of the dry lake bed itself in the future — creating visual prompts that help people take pleasure in a huge, managed landscape that comprises pools, sand and mudflats as well as dust-control gravel raked into sinuous curves. To many, it will be much less familiar than, say, a white-sand beach or lake front.

And Robinson concedes that amplifying the beauty of such a landscape will not be simple. Unlike people working on many environmental design projects, architects and artists at Owens Lake must balance the aesthetics of the landscape with the need to conserve wildlife and water, and stave off further toxic dust storms.

Robinson ends the book with tools that might help landscape architects — and the public — to see the future of Owens Lake as an amenity. Beyond making it non-toxic and ecologically sound, the city could call the project a success if the public champions

“Massive human developments need not always result in decimation.”

this reinvented land. To this end, one of the tools Robinson introduces is a robot that rakes sandpits with various designs and connects with 3D-modelling software that renders how the lake bed might look in various scenarios. These parameters can be drawn through the reams of data already collected on factors such as water flow, ecology and cost. To encourage public participation, Robinson has created an arcade-style game in which users explore how changes in design alter dust, water and habitat on the basis of computer models. They then print a postcard from their imagined territory.

At times, *The Spoils of Dust* is a dense read. But its gorgeous maps, graphs and photographs celebrate a landscape that others might dismiss as post-apocalyptic. Robinson makes a convincing case that massive human developments need not always result in decimation. "Even more improbable than the control of the lake's fearsome dust storms," he writes, "is the fact of its strange rebirth." ■

Amy Maxmen is a senior reporter for *Nature*, based in San Francisco, California.

Correspondence

Networks, mentors and role models

Women in the Middle East and North Africa (MENA) have fewer career opportunities in the health and science sectors than do men, mainly because of cultural norms and traditional obligations. In the absence of the legal equity afforded to some female scientists in the West, I suggest that mentoring, networking opportunities and inspirational role models can help (see also *Nature* **560**, 164; 2018).

The region's limited research opportunities are generally offered to men (see go.nature.com/2p4ruxt). And family expectations lead to a high drop-out rate for female researchers (see go.nature.com/2zget46).

As the Three Circles of Alemat initiative (<https://tca.jssr.jo/>) shows, supportive mentors can help female scientists to balance family and work and to progress along career paths that are conventionally dominated by men. Other professional women can act as role models. And male scientists need to understand that family responsibilities should no longer concern only women; they, too, must stand up for equality.

After all, it was an Arab Muslim woman, Fatima al-Fihri, who founded the world's first university, in the ninth century.

Hossein Bannazadeh Baghi
Tabriz University of Medical Sciences, Tabriz, Iran.
hbannazadeh@tbzmed.ac.ir

Pay PIs' salaries and benefits

Principal investigators (PIs) working in basic research at US biomedical institutions have to draw upwards of 65% of their own salary and benefits from the direct costs covered by their research grants. The situation discourages top talent and is a roadblock to diversity.

Yet institutions continue to take on PIs because each is a new source of grants and funds to cover indirect costs. PIs therefore

need to spend more and more time applying (and reapplying) for grants. Up to US\$6 billion of the annual research funding paid out by the National Institutes of Health is spent on PIs' salaries and benefits (L. R. Pool *et al.* *FASEB J.* **30**, 1023–1036; 2016). Government institutions pay the salaries and benefits of teachers and firefighters — research centres should do the same for PIs.

US public funding for universities is falling. Increasing tuition fees is not an acceptable option. Instead, funds to cover faculty salaries could continue to come mainly from federal agencies — but through direct negotiation with institutions, in the same way that indirect costs are met. Philanthropic funds should be directed more towards programmes that include salaries and less towards new buildings. A longer-term solution might be to both eliminate the tenure system and plan for a sustainable faculty pool that is based on merit.

Research institutions rightly opposed US President Donald Trump's proposal last year to slash coverage for indirect costs. However, it is the same institutions that benefit from and perpetuate the status quo in research-faculty employment.

Emily Bernstein *Icahn School of Medicine at Mount Sinai, New York, New York, USA.*

Alexander Meissner *Max Planck Institute for Molecular Genetics, Berlin, Germany.*

Miguel Ramalho-Santos
Lunenfeld-Tanenbaum Research Institute, University of Toronto, Canada.
mrsantos@lunenfeld.ca

Lab agreements improve mentoring

We suggest that written lab agreements on best practices help to improve mentoring of students and trainees (see also *Nature* **561**, 7; 2018).

Such agreements focus on the responsibilities of mentor and trainee, on facilitating

communication between them and on their mutual expectations in matters including availability, contributions to lab life, tasks and meetings (see also D. Norris *et al.* *Nature* **557**, 302–304; 2018). They promote collaborative discussion and accountability, and should be updated regularly to incorporate feedback.

Given the potential importance of such agreements for lab members, faculty members should be recognized for creating them. The more scientists involved in the endeavour, the more resources, dialogue and momentum there will be.

June Gruber* *University of Colorado Boulder, USA.*
june.gruber@colorado.edu

*On behalf of 5 correspondents (see go.nature.com/2jfpcz7 for details).

we can eat. That means raising them on extensive grasslands, rotating integrated crop-livestock systems and using by-products, unavoidable food waste and crop residues as feed.

Feeding animals exclusively on such materials would greatly reduce the availability and hence the consumption of meat and dairy products, as well as the use of water, energy and pesticides — thereby cutting greenhouse-gas emissions.

Philip Lymbery* *University of Winchester, UK.*

*On behalf of 58 co-signatories (see go.nature.com/2z32kkn for full list).

philip@ciwf.org

3D print unique specimens

Museum collections should use 3D printing to create replicas of their most important specimens. This would guard against loss or damage, as has occurred on a massive scale in Brazil. Two of South America's largest scientific collections have burned down: the National Museum in Rio de Janeiro, in September, and the Butantan Institute in São Paulo in 2010. Copies would also obviate the need for direct access to specimens, which is an advantage for distant scholars.

When I analysed collected material for my master's thesis on systematics 10 years ago, I measured more than 600 individuals of *Scinax granulatus*, an amphibian from South America. As a Brazilian living back home at the time, I was unable to measure the most important specimen — the name-bearing holotype — because it is in a German museum. Indeed, much of the Southern Hemisphere's biodiversity was described from specimens held in Europe and North America.

3D printed replicas would benefit taxonomy and systematics worldwide.

Luis Fernando Marin da Fonte
Trier University, Trier, Germany.
pulchella@gmail.com

Kuen Charles Kao

(1933–2018)

Engineer who proposed optical fibre communications that underpin the Internet.

Today, fibre-optic cables carry more than 95% of all digital data around the world, underpinning the Internet. In 1966, it was Kuen Charles Kao (Charlie to his colleagues) who proposed the use of optical fibres as a universal medium for communication, and calculated how it might be done. Given the rudimentary technology available at the time, it was a leap of imagination, bordering on science fiction. For this work, Kao won a share of the Nobel Prize in Physics in 2009.

Kao was born on 4 November 1933 into Shanghai high society, to an academic lawyer father and poet mother. Introverted and geeky, Kao was educated at home with his younger brother Timothy before going to French- and English-speaking schools. In 1953, he moved to England to study at Woolwich Polytechnic (now the University of Greenwich in London).

Graduating in electrical engineering in 1957, he joined Standard Telephones and Cables, part of the conglomerate International Telephone & Telegraph (ITT). There he met his wife, fellow engineer Gwen Mae-wan Wong. He turned down a lectureship at Loughborough Polytechnic, UK, to do an industrial PhD in the company's research arm — the Standard Telecommunication Laboratories (STL) in Harlow, UK. Similar to Bell Labs in the United States (although less well funded), STL was a nursery for future academic and industrial leaders, heady with creativity, camaraderie and resourcefulness. Kao joined the group of Toni Karbowiak, working alongside another British telecommunications pioneer, Alec Reeves.

At the time, telecommunications used coaxial electronic cables or broadcast radio signals in the megahertz frequency range. Growing demand for information transfer meant moving to higher, microwave frequencies (gigahertz), with major research programmes set up around the world to find a way to guide signals from source to destination. The front-runner technology was hollow metal waveguides, pioneered in the 1950s by Harold Barlow, Kao's external PhD supervisor at University College London. Costly and impractical, these metal tubes needed to be laid in straight lines. Karbowiak, a seasoned microwave engineer and former PhD student of Barlow, knew that new ideas were needed.

In the early 1960s, just as the laser came about, Karbowiak asked Kao to look at an optical analogue of a microwave waveguide. Optical signals have an even higher frequency (hundreds of terahertz), and so can carry



more information. The idea of making a waveguide for the transmission of light over hundreds of kilometres was breathtaking. It meant shrinking the waveguide from a few centimetres across to something as thin as a human hair, just 100 or so micrometres wide. Glass was the most optically transparent material known, and had the advantages of being potentially flexible and resistant to lightning. But could it be made pure and clear enough? George Hockham, a talented young theorist, was assigned to help Kao.

They started pragmatically; given the power available from the earliest lasers of the time, the sensitivity of detectors, and the distance between UK telecommunications switching centres, they calculated that a signal could afford to lose only 20 decibels (a logarithmic measure of power) per kilometre travelled — equivalent to a 99% power loss after 1 km. This was an ambitious target: the best glasses at the time had losses some 10^{98} times greater, of around $1,000 \text{ dB km}^{-1}$. Kao systematically analysed the absorption, reflection and scattering of different glasses, while Hockham did waveguide-dimension calculations. Their landmark 1966 paper concluded that the task, although difficult, was theoretically possible (K. C. Kao and G. A. Hockham *Proc. Inst. Electr. Eng.* **113**, 1151–1158; 1966).

The paper went almost unnoticed, except at the research labs of the UK General Post Office (the telecommunications arm of which later became British Telecom, now BT) and the Ministry of Defence. Both organizations set up research programmes in this area, attracted by the idea of a lower-cost alternative to microwave waveguides.

But there was much scepticism — the gap between theory and practice was huge. To

convince others, Kao measured the losses in the purest glasses he could find, now aided by Mervin Jones (Hockham left to start his own antenna-technology research group in 1967). They devised a complex and elegant set-up to measure very low values of loss in rods of fused silica glass about the length of a ruler. They published their results in 1969 (M. W. Jones and K. C. Kao *J. Phys. E Sci. Instrum.* **2**, 331; 1969). The following year, Robert Maurer's group at US firm Corning Glass broke the 20 dB km^{-1} limit in optical fibres of around 1 km long. Together with reports of the first continuous-wave room-temperature semiconductor laser in 1970, this convinced the doubters, sparking research efforts worldwide.

The optical fibre revolution had begun. Much of the work was done at STL and at the Post Office research labs in Britain, in fierce competition with Bell Labs and the US telecommunications firm AT&T. In 1977, the UK Post Office was the first to install optical fibres in its telecommunications network. The first transatlantic system followed in 1988.

From 1970 to 1974, Kao set up the electrical engineering department at the Chinese University of Hong Kong (CUHK), returning to STL in the holidays to keep abreast of research. In 1974, Kao went to work for ITT in the United States, where he rose to director of corporate research in 1985. In 1986, he returned to CUHK as its vice-chancellor, where, for nine years, he used his connections to strengthen the university's research base and make it internationally competitive.

In the mid-2000s, Kao developed Alzheimer's disease. He attended the 2009 Nobel award ceremony and celebrations afterwards, always bearing a smile, but his Nobel speech was read by his wife Gwen. He died in Hong Kong on 23 September.

Kao's legacy is hard to overestimate. Today, his 1966 predictions have been exceeded by six orders of magnitude, with fibre losses of less than 0.15 dB km^{-1} . Kao's determination inspired those of us who worked at STL right up to its closure in 2009. The site, now a technology business hub, is named Kao Park in honour of its most famous resident. ■

Polina Bayvel is professor of optical communications and networks at University College London. She worked at STC/STL in the 1990s and co-organized the Royal Academy of Engineering's 2010 celebrations marking Kao's Nobel prize and 50 years of the laser.

e-mail: p.bayvel@ucl.ac.uk

FORUM Regeneration

Defining adult stem cells

Adult tissues must maintain themselves and regenerate after damage. But are these crucial functions mediated by dedicated populations of stem cells, or do differentiated cells adopt stem-cell-like properties according to an organ's needs? Here, two scientists present evidence from both sides of the debate.

Dedicated to the job

PURA MUÑOZ-CÁNOVES

The term stem cell was coined at the end of the nineteenth century to propose the notion of a common progenitor cell for distinct blood lineages^{1,2}. The existence of this progenitor, called a haematopoietic stem cell (HSC), was finally proved in the 1960s³. The discovery of HSCs led to the defining concept of a stem cell as a self-renewing cell positioned at the top of a hierarchy, giving rise to a range of fully differentiated, specialized cell types at the end of the hierarchy's branches. This type of dedicated adult stem cell has since been identified in several tissues.

A second clear example of a dedicated stem-cell population is the satellite cells of skeletal muscle⁴. There are many parallels between these cells and HSCs. Both reside in specialized, protective niches — HSCs in the bone marrow and satellite cells in bundles of muscle fibres (myofibres). The niche enables both cell types to exist in a dormant state until needed, dividing as little as possible to minimize the risk of accumulating harmful genetic mutations. And, like HSCs, satellite cells are activated and divide in response to damage, subsequently self-renewing and differentiating into newly regenerated myofibres along a unidirectional, hierarchical pathway⁵ (Fig. 1a).

HSCs were first identified through experiments demonstrating that the bone marrow could repopulate the blood system of mice whose own marrow had been destroyed³. Likewise, cell-tracing studies and experiments in which satellite cells were grafted into damaged muscle have shown that myofibre repair involves the direct participation of satellite cells. Furthermore, mice genetically depleted of satellite cells lack the capacity to form new myofibres, confirming satellite cells as genuine adult stem cells (reviewed in ref. 5).

But although attempts to find such rare, 'professional' stem cells have been successful in some tissues, in others, stem-cell-like processes can be more varied. Indeed, it is

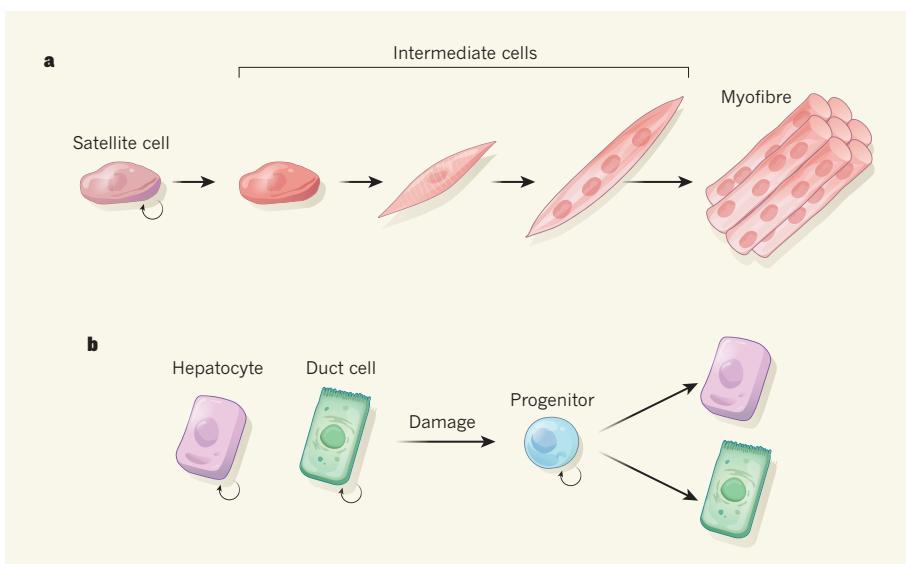


Figure 1 | Professional and facultative stem cells. **a**, Satellite cells are a dedicated (professional) population of muscle stem cells. Under normal conditions (in homeostasis), satellite cells are dormant (not shown). Following muscle damage, the stem cells begin both to self-renew (curved arrow) and to give rise to a series of intermediate progeny. The differentiation cascade terminates with the formation of fully differentiated, mature muscle cells called myofibres, which contain multiple nuclei. **b**, By contrast, the liver contains no known professional stem cells. Under homeostasis, progenitor cells for both the liver's main cell type (hepatocytes) and bile-duct cells maintain their own populations by proliferating. Following damage, these unipotent progenitors can also acquire a bi-potential progenitor state (here shown for the duct cell), from which they can self-renew and give rise to both hepatocytes and duct cells. Whether a bi-potential progenitor exists in homeostasis is yet to be confirmed (not shown).

becoming clear that, in some cases, repair can involve regression of differentiated cells into a less-differentiated state from which they repopulate the tissue. This is in stark contrast

to the situation in blood and skeletal muscle; dedifferentiation of other niche cell types cannot compensate for HSC or satellite-cell loss^{6,7}.

The lack of obvious physical stem-cell populations in some tissues has prompted increasingly strident challenges to the definition of adult stem cells as discrete entities that follow unidirectional hierarchies, and has led to calls for an emphasis on the more diverse,

plastic properties of stem cells. But to shift the focus away from professional stem cells risks negating the benefits of identifying and understanding these dedicated populations.

The ability to use professional stem cells for grafting experiments makes the cells easier to harness for therapies and experiments than more-plastic stem-cell-like populations. Indeed, HSC transplantation is increasingly used to treat a range of diseases, including blood, metabolic and immunological disorders and some cancers⁸. Satellite-cell transplants are a promising tool for the treatment of muscle diseases, particularly those associated with reduced numbers of satellite cells and impaired regenerative capacity, such as ageing-associated and inherited muscle disorders⁹. In the midst of calls to expand the definition of stem cells, we should remember

The ability to use professional stem cells for grafting experiments makes the cells easier to harness for therapies.

that as-yet-unknown, dedicated stem-cell populations might still await discovery. Their identification could have major clinical implications. ■

Pura Muñoz-Cánoves is in the Department of Experimental and Health Sciences, Pompeu Fabra University and ICREA, 08003 Barcelona, and the Spanish National Centre for Cardiovascular Research, Madrid, Spain. e-mail: pura.munoz@upf.edu

Regeneration on call

MERITXELL HUCH

Unlike blood and muscle stem cells, which reside in protected niches, epithelial tissues that line or bud off from the body's tubes are often exposed to external or internal stressors. An HSC-like branching hierarchy in which a single progenitor sits atop a direct line of descendants seems a very unsafe evolutionary solution for this type of tissue — dependence on a single 'master' cell would put the tissue at risk of disintegration should that cell type die. An alternative approach involving overlapping hierarchies with two or more entry points seems a more secure means of solving the problem. This idea suggests that facultative stem cells, which can act as stem cells if needed, but do not always do so, must exist.

The debate about whether the hierarchical HSC-like model fits other systems¹⁰ has been influenced by the tendency of researchers to consider normal organ maintenance (homeostasis) as equivalent to regeneration and repair, despite the highly divergent intrinsic cellular responses involved in the two phenomena. Repair often requires a higher level of proliferation than does homeostasis — therefore, bone fide stem cells that can mediate homeostasis cannot always repopulate a damaged tissue. This is where facultative stem cells come in.

One example of this phenomenon can be found in the intestinal epithelium, which is highly proliferative both in homeostasis and following injury. A population of dedicated stem cells maintains this tissue under normal conditions. These are known as crypt-base columnar cells, and they self-renew and differentiate into several cell types¹¹. However, if the tissue is injured or the stem-cell population depleted, non-proliferative cells that have begun to differentiate or have even fully matured can revert to a stem-cell-like state to help repopulate the tissue¹¹. Thus, cellular plasticity is key to gut maintenance in different conditions.

Unlike the intestine, most tissues undergo cellular turnover only slowly in everyday life, and show an increased proliferative capacity

that enables them to repair some (but not all) structures following injury. However, a few tissues that typically have low turnover, including the liver and lung, can completely regenerate following injury. The cells that enable this remarkable response have been extensively investigated, and have provided further examples of facultative stem cells.

The lung, like the intestine, has a population of true 'HSC-like' stem cells that maintain the airway by means of homeostasis. Following injury, mature differentiated cells called club cells can dedifferentiate and behave as facultative stem cells^{12,13}. By contrast, the existence of any dedicated stem cell in the liver has yet to be confirmed. During homeostasis, two liver-cell types, hepatocytes and ductal cells, seem to maintain their respective cell types through proliferation. But following damage, at least in zebrafish¹⁴ and mice¹⁵, facultative stem cells arise from differentiated cells called cholangiocytes. In mice, cholangiocytes revert to a bi-potent stem-cell-like state that facilitates the regeneration of both hepatocytes and ductal cells¹⁵ (Fig. 1b).

These three examples highlight ways in which different organs have solved similar problems. That brings to mind the natural-selection pressures that lead different groups of animals to achieve various solutions to common habitat challenges — developing different strategies to combat the extreme cold weather at the poles, for instance. It is tempting to speculate that the battle to maintain tissues in a demanding environment that involves constant turnover and exposure to damage has resulted in the existence of a range of back-up strategies through

which facultative stem cells help to ensure tissue integrity. A definition of stem cells that encompasses the existence of the full range of these plastic cell types is essential if we are to truly understand the nature of regeneration. ■

Meritxell Huch is at the Gurdon Institute, University of Cambridge, Cambridge CB2 1QN, UK, and at the Wellcome Trust-Medical Research Council Stem Cell Institute, Cambridge. e-mail: m.huch@gurdon.cam.ac.uk

1. Pappenheim, A. *Virchows Arch. Eur. J. Pathol.* **145**, 587–643 (in German) (1896).
2. Ramalho-Santos, M. & Willenbring, H. *Cell Stem Cell* **1**, 35–38 (2007).
3. Becker, A. J., McCulloch, E. A. & Till, J. E. *Nature* **197**, 452–454 (1963).
4. Mauro, A. *J. Biophys. Biochem. Cytol.* **9**, 493–495 (1961).
5. Relaix, F. & Zammit, P. S. *Development* **139**, 2845–2856 (2012).
6. Lepper, C., Partridge, T. A. & Fam, C.-M. *Development* **138**, 3639–3646 (2011).
7. Sambasivan, R. *et al.* *Development* **138**, 3647–3656 (2011).
8. Chivu-Economescu, M. & Rubach, M. *Curr. Stem Cell Res. Ther.* **12**, 124–133 (2017).
9. Pini, V., Morgan, J. E., Muntoni, F. & O'Neill, H. C. *Curr. Stem Cell Rep.* **3**, 137–148 (2017).
10. Clevers, H. & Watt, F. M. *Annu. Rev. Biochem.* **87**, 1015–1027 (2018).
11. Tetteh, P. W., Farin, H. F. & Clevers, H. *Trends Cell Biol.* **25**, 100–108 (2015).
12. Tata, P. R. *et al.* *Nature* **503**, 218–223 (2013).
13. Rawlins, E. L. *et al.* *Cell Stem Cell* **4**, 525–534 (2009).
14. Choi, T.-Y., Ninov, N., Stainier, D. Y. R. & Shin, D. *Gastroenterology* **146**, 776–788 (2014).
15. Raven, A. *et al.* *Nature* **547**, 350–354 (2017).

This article was published online on 29 October 2018.

ASTRONOMY

A key piece in the exoplanet puzzle

The detection of a low-mass exoplanet on a relatively wide orbit has implications for models of planetary formation and evolution, and could open the door to a new era of exoplanet characterization. SEE LETTER P.365

RODRIGO F. DÍAZ

For decades, astronomers have looked for planets around a nearby star known as Barnard's star. On page 365, Ribas *et al.*¹ report evidence for such a planet, based on more than 20 years of data. Detailed information about the planet could be revealed by the next generation of astronomical instruments.

Planets around stars other than the Sun are known as exoplanets. They are extremely faint compared with their host stars, and their orbits are typically too small to be resolved — even

using the largest telescopes available today. As a result, the latest high-resolution imaging techniques are limited to giant planets on wide orbits around nearby stars^{2,3}.

Most of what is currently known about the properties, formation and evolution of exoplanets therefore comes from indirect methods that measure variations in the light received on Earth from host stars. One of the most fruitful of these methods, used by Ribas and colleagues, is the radial-velocity technique. It involves measuring changes in the velocity of a host star along the line-of-sight of an observer,

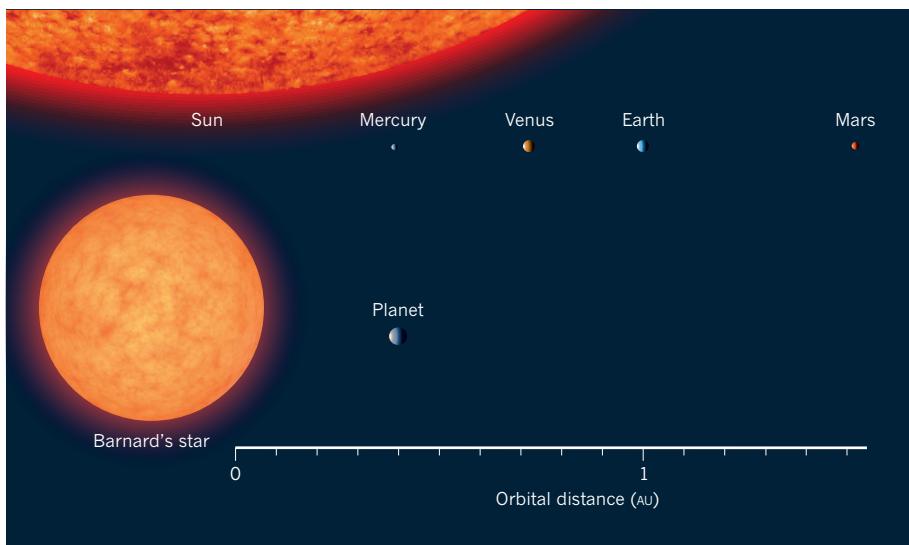


Figure 1 | A planet around Barnard's star. Ribas *et al.*¹ report evidence that a low-mass planet orbits a nearby star called Barnard's star. Shown here is the orbital distance of the planet, compared with those of the four inner planets of the Solar System. The distances are given in astronomical units (1 AU is the average separation between Earth and the Sun). The discovered planet is on a relatively wide orbit. The sizes of all the objects are approximately to scale.

and is sensitive to the mass of the exoplanet. However, because the measurements depend on an unknown value (the inclination of the planet's orbit), the technique provides only a lower bound on the planet's mass⁴.

Using the radial-velocity method to detect planets on long-period orbits is difficult, because the larger the orbital distance, the smaller the planetary signal. The transit method, which measures the drop in brightness as a planet passes in front of its host star, is also ineffective for planets on long-period orbits. This limitation has restricted the detection and characterization of exoplanets mostly to close-in companions, especially in the case of low-mass planets, whose signals are small.

Ribas and colleagues' announcement of a planet around Barnard's star with an orbital period of 233 days and a mass of at least 3.2 times that of Earth pushes the limits of the radial-velocity technique. Barnard's star belongs to a family of stars known as M dwarfs, which are cooler and much less massive than the Sun. M dwarfs are prime targets for planetary searches, because they favour the detection of small companions.

The orbital distance of the reported planet is similar to that of Mercury from the Sun (Fig. 1). This places the planet close to the snow line of Barnard's star — the region out from the star beyond which volatile elements can condense. The snow line is a key region of planetary systems. In particular, there are indications that the building blocks of planets are formed there⁵.

It is currently thought that these building blocks grow by collecting material from their surroundings to become planetary cores and then fully fledged planets as they migrate towards their host stars⁶. Until now, only giant planets had been detected at such a distance

from their stars. The authors' discovery of a low-mass planet near the snow line places strong constraints on formation models for this type of planet.

Ribas *et al.* report that Barnard's star seems to be devoid of close-in companions. In particular, the authors put stringent constraints on the presence of planets in the habitable zone around the star — the region in which liquid water could exist on the surface of a rocky planet. However, they do provide an unconfirmed hint of a planet farther away from the star than the detected planet.

What makes the authors' discovery even more remarkable is that Barnard's star is only 1.8 parsecs (less than 6 light years) away from the Sun⁷. This makes the planetary system the closest single-star system to the Sun. The Alpha Centauri triple-star system is the only system that is closer, and also hosts at least one low-mass planet⁸.

Barnard's star has been monitored for more than 20 years. Ribas and co-workers used hundreds of radial-velocity observations that were obtained with different instruments by many different projects and researchers. These measurements were crowned by an intense observing campaign with the CARMENES spectrograph⁹, which is located at the Calar Alto Observatory in Spain.

In their analysis, the authors had to be particularly careful in accounting for stellar activity. For many years, exoplanet researchers have been struggling with the effects of stellar activity — for example, rotating stellar spots and active regions, and long-term activity cycles similar to the 11-year cycle of the Sun. These phenomena can easily mimic the effects of planetary companions¹⁰, especially in the case of low-amplitude signals^{11,12} such as the one detected by Ribas and colleagues.

The authors used a few different methods to account for the effects of stellar activity, and to check whether the planetary detection depended on how such effects were corrected for. One of these methods resulted in a drastic reduction in the statistical significance of the detection, and therefore casts doubt on the discovery. However, this method is prone to false negatives¹³ and, using simulations, the authors showed that their detection can be reproduced by stellar activity in only 0.8% of cases.

Difficult detections such as this one warrant confirmation by independent methods and research groups. Amassing an independent radial-velocity data set to confirm the existence of the planet seems unfeasible in the near future, but the closeness of Barnard's star to the Sun means that confirmation should be possible through other means. For example, a signal for the planet might be detectable in astrometric data — precision measurements of stellar positions — from the Gaia space observatory that are expected to be released in the 2020s. Such a signal would confirm the presence of the planet, reveal the planet's actual mass (as opposed to a lower bound on the mass) and provide complementary information on the planet's orbit.

Even more excitingly, the next generation of ground-based instrumentation, also coming into operation in the 2020s, should be able to directly image the reported planet, and measure its light spectrum. Using this spectrum, the characteristics of the planet's atmosphere — such as its winds and rotation rate — could be inferred. This remarkable planet therefore gives us a key piece in the puzzle of planetary formation and evolution, and might be among the first low-mass exoplanets whose atmospheres are probed in detail. ■

Rodrigo F. Díaz is at the Institute of Astronomy and Space Physics, National Council of Scientific and Technical Research and University of Buenos Aires, 1428 Buenos Aires, Argentina.

e-mail: rodrigo@iafe.uba.ar

1. Ribas, I. *et al.* *Nature* **563**, 365–368 (2018).
2. Wagner, K. *et al.* *Science* **353**, 673–678 (2016).
3. Vigan, A. *et al.* *Mon. Not. R. Astron. Soc.* **407**, 71–82 (2010).
4. Murray, C. D. & Correia, A. C. M. in *Exoplanets* (ed. Seager, S.) 15–23 (Univ. Arizona Press, 2010).
5. Drażkowska, J. & Alibert, Y. *Astron. Astrophys.* **608**, A92 (2017).
6. Ida, S. & Lin, D. N. C. *Astrophys. J.* **719**, 810–830 (2010).
7. Gaia Collaboration. *Astron. Astrophys.* **616**, A1 (2018).
8. Anglada-Escudé, G. *et al.* *Nature* **536**, 437–440 (2016).
9. Trifonov, T. *et al.* *Astron. Astrophys.* **609**, A117 (2018).
10. Queloz, D. *et al.* *Astron. Astrophys.* **379**, 279–287 (2001).
11. Feroz, F. & Hobson, M. P. *Mon. Not. R. Astron. Soc.* **437**, 3540–3549 (2014).
12. Robertson, P., Mahadevan, S., Endl, M. & Roy, A. *Science* **345**, 440–444 (2014).
13. Feng, F., Tuomi, M., Jones, H. R. A., Butler, R. P. & Vogt, S. *Mon. Not. R. Astron. Soc.* **461**, 2440–2452 (2016).

Gut microbes alter fly walking activity

A gut bacterium has been found to modulate locomotor activity in the fruit fly *Drosophila melanogaster*. This effect is mediated by the level of a sugar and the activity of neurons that produce the molecule octopamine. SEE LETTER P.402

ANGELA E. DOUGLAS

Is the refrain that “my microbes make me do it” true? Scientific reviews and the popular press often report that microbes can influence many aspects of the behaviour of their healthy animal or human hosts, from cognition to social interactions to emotional state¹. If this is true, perhaps future microbial-based therapies might be used to improve mental health. Yet the evidence for most of these claims of microbial effects is limited. Experiments are urgently needed that definitively test whether bacteria can have a causal role in behaviour, and that identify the underlying mechanisms. On page 402, Schretter *et al.*² provide a superb example of rigorous scientific analysis, in which they demonstrate that the walking activity of the fruit fly *Drosophila melanogaster* can be affected by a specific gut bacterium. The authors identify a bacterial enzyme that mediates this effect, and establish aspects of the mechanism by which the fruit fly responds to the bacterium.

Schretter *et al.* used standard techniques to analyse the bacterial residents of the gut. The authors compared walking activity in flies harbouring their natural gut microbes (the microbiota) and flies that had been treated to eliminate the gut bacteria, and they observed that the treated flies were hyperactive in comparison to the others. These hyperactive flies walked faster and for longer than the others, but their daily (circadian) rhythms of activity and sleep were not perturbed. To determine the microbes associated with this effect, Schretter *et al.* supplied the hyperactive flies with various bacteria, and found that the bacterium *Lactobacillus brevis* restored walking activity to the level observed in flies that retained their complete microbiota.

There is a common expectation that gut microbes influence animal behaviour by producing small metabolites, including neurotransmitter molecules, which interact directly with the nervous system in the gut or that enter the bloodstream and from there reach the brain^{3,4}. However, the bacterial product identified by Schretter and colleagues as involved in walking behaviour does not fit this paradigm. The authors provide persuasive evidence that the presence of the sugar-modifying enzyme xylose isomerase, which is produced

by *L. brevis*, reduces locomotor activity in *D. melanogaster* (Fig. 1). Further experiments revealed that supplying xylose isomerase to flies whose bacteria had been eliminated was necessary and sufficient to modulate fly locomotion.

How does xylose isomerase cause *D. melanogaster* to slow down? The enzyme mediates the interconversion of certain sugar molecules — the change of glucose into fructose, for example. Schretter and colleagues found that the flies treated to remove their gut bacteria have a higher level of the sugar trehalose than do those that retain their usual microbiota. Perhaps this means that xylose isomerase decreases the availability of a glucose substrate needed for the synthesis of trehalose. The authors administered trehalose to flies that lacked gut bacteria and had been provided

with xylose isomerase, and report that the trehalose treatment caused the flies’ walking speed to increase.

The authors proceeded to investigate the neural basis of the hyperactivity phenomenon. They used genetic approaches to activate neurons that regulate locomotion in *D. melanogaster* and the results focused their attention on a type of neuron called an octopaminergic neuron, which produces the neurotransmitter molecule octopamine. Schretter *et al.* found that the walking activity of flies that lacked gut bacteria but had been given xylose isomerase was increased by activation of the genes encoding enzymes needed for the synthesis of octopamine. Such an effect on locomotion was not observed for other neurotransmitters they tested. Furthermore, the authors observed that the flies with their natural microbiota and those that had been treated to remove gut bacteria but had received xylose isomerase both walked faster if they received octopamine.

Octopamine is a well-characterized regulator of locomotion in flies. In vertebrates, the neurotransmitter molecule noradrenaline, which is related in structure to octopamine, fulfils a similar role in promoting physical activity⁵. Work remains to be done to fill in the gaps in explaining how xylose isomerase affects the level of trehalose in the fruit fly and the activity of octopamine-producing neurons in the brain. Nevertheless, one key conclusion emerges: the effect of bacterial products on fly

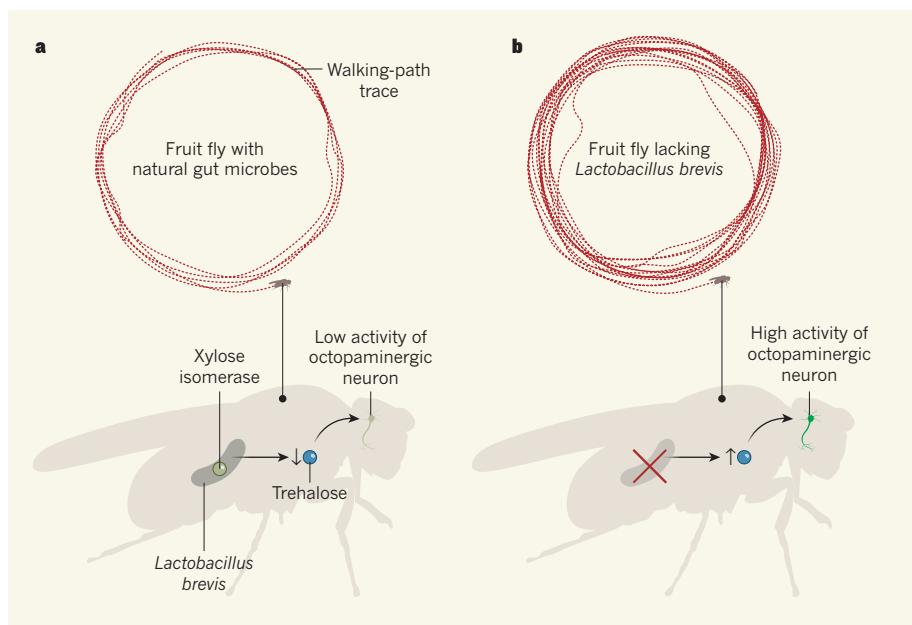


Figure 1 | A gut bacterium affects walking activity in the fruit fly *Drosophila melanogaster*. Schretter *et al.*² used imaging approaches to track fly movement and found that (a) fruit flies that have their natural gut microbes were less active than (b) flies that lack the gut bacterium *Lactobacillus brevis*. **a.** The authors reveal that the enzyme xylose isomerase produced by *L. brevis* is key to this phenomenon. This enzyme modifies certain sugars, which leads, by an unknown process, to a decrease in the level of the sugar trehalose in the body of flies in which this bacterial enzyme is present. The results of the authors’ experiments are consistent with a model in which a decrease in trehalose is accompanied by a decrease in the activity of octopaminergic neurons (those that produce the neurotransmitter molecule octopamine) that regulate fly locomotion. **b.** Compared with flies that have their natural gut microbes, flies lacking *L. brevis* have higher levels of the sugar trehalose in their body and are proposed to have higher activity of octopaminergic neurons.

locomotion is mediated by the modulation of known circuits that control behaviour and not through previously unknown regulatory mechanisms.

Why does *L. brevis* make xylose isomerase? It should not be assumed that this is a specific adaptation to life in a *D. melanogaster* host. This bacterium is not specialized to exist only in the fruit fly gut. It maintains substantial free-living populations⁶ and is neither universally present nor abundant in *D. melanogaster* populations in the natural environment⁷. Xylose isomerase probably functions to increase the diversity of the carbon sources that *L. brevis* can exploit, as is the case for the many other bacteria that produce this enzyme. It would be interesting to learn the outcome of experiments comparing the abundance in *D. melanogaster* of resident wild-type

L. brevis and of *L. brevis* mutants lacking xylose isomerase, to determine whether this enzyme enhances the fitness of the bacterium and whether any fitness effects depend on fly locomotor activity. ■

The most important question to ask next is whether the effect of *L. brevis* and xylose isomerase on the locomotor activity of *D. melanogaster* is relevant to animal behaviour in general, including that of humans and other mammals. As with many other discoveries first made in *D. melanogaster*⁸, perfect correspondence with mammalian systems is unlikely. Schretter and colleagues' study does, however, alert microbiologists and those studying animal behaviour to pay attention to the enzymes of gut bacteria and their possible effects on sugar metabolism and on the neuronal circuits regulating walking activity. ■

Angela E. Douglas is in the Departments of Entomology and of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA.

e-mail: aes326@cornell.edu

1. Mohajeri, M. H., La Fata, G., Steinert, R. E. & Weber, P. *Nutr. Rev.* **76**, 481–496 (2018).

2. Schretter, C. E. et al. *Nature* **563**, 402–406 (2018).

3. Cryan, J. F. & Dinan, T. G. *Nature Rev. Neurosci.* **13**, 701–712 (2012).

4. Sharon, G. et al. *Cell Metab.* **20**, 719–730 (2014).

5. Roeder, T. *Annu. Rev. Entomol.* **50**, 447–477 (2005).

6. Duar, R. M. et al. *FEMS Microbiol. Rev.* **41**, S27–S48 (2017).

7. Bost, A. et al. *Mol. Ecol.* **27**, 1848–1859 (2018).

8. Letsou, A. & Bohmann, D. *Dev. Dyn.* **232**, 526–528 (2005).

This article was published online on 24 October 2018.

METABOLISM

A back door to improved health

The coenzyme NAD⁺ can be produced from the amino acid tryptophan. It emerges that inhibiting an enzyme that degrades an intermediate in this pathway can help to combat kidney and liver diseases in mouse models. SEE ARTICLE P.354

SAMIR M. PARIKH

Throughout the history of life on Earth, there has been a requirement for small molecules called nucleotides. Long chains of nucleotides make up the genetic code, and single nucleotides transduce signals or transfer energy. In addition, a dimeric form of nucleotide called nicotinamide adenine dinucleotide (NAD⁺) serves at least two pivotal cellular functions. The first is to shuttle high-energy electrons to enzymatic complexes found in organelles called mitochondria, where their energy can be efficiently harvested; the second is as a substrate for enzymes such as sirtuins, which regulate many cellular behaviours. On page 354, Katsyuba *et al.*¹ shed light on a fundamental mechanism by which the correct levels of NAD⁺ are maintained in cells, and demonstrate how augmenting this pathway can affect disease.

In simple terms, the available pool of NAD⁺ in a cell is governed by the balance between its generation and its consumption. The predominant pathway by which NAD⁺ is generated in

rodents relies on the recycling of a molecule called nicotinamide (Nam) that is either ingested or released by enzymes that consume NAD⁺ (Fig. 1). There are several other routes of NAD⁺ production, including a *de novo* synthesis pathway that starts with

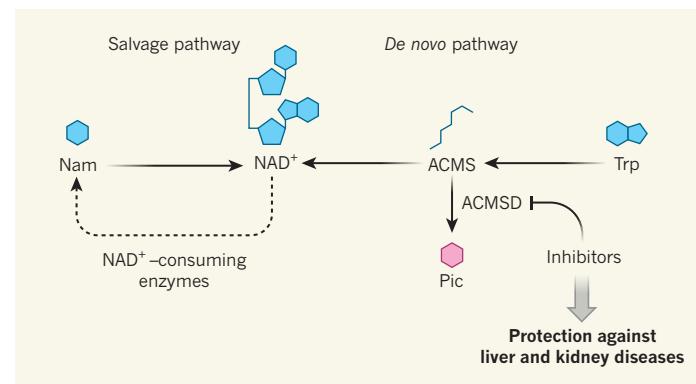


Figure 1 | NAD⁺ biosynthesis in disease. When the coenzyme nicotinamide adenine dinucleotide (NAD⁺) is consumed by enzymes, nicotinamide (Nam) is generated as a reaction product. Through a recycling mechanism called the salvage pathway, NAD⁺ can then be regenerated. Nam salvage is considered the predominant mechanism for NAD⁺ biosynthesis, but NAD⁺ can also be generated through multiple other routes. One of these is the *de novo* pathway, whereby the amino acid tryptophan (Trp) is converted to NAD⁺ through several intermediates, including α -amino- β -carboxymuconate- ϵ -semialdehyde (ACMS). This pathway can be depleted by the enzyme ACMS decarboxylase (ACMSD), which degrades ACMS to picolinic acid (Pic). Katsyuba *et al.*¹ report that chemical inhibition of ACMSD raises NAD⁺ levels in mice and nematode worms, and improves outcomes in mouse models of liver and kidney diseases.

the essential amino acid tryptophan (Trp)². Mutations that disrupt the enzymes responsible for converting Trp to NAD⁺ result in multi-system developmental alterations in humans³, demonstrating the importance of this *de novo* pathway.

Katsyuba *et al.* set out to study α -amino- β -carboxymuconate- ϵ -semialdehyde (ACMS), an unstable and little-studied intermediate of the Trp pathway. ACMS can either spontaneously convert to the next intermediate on the path to NAD⁺, or can be degraded by a train of enzymes, starting with ACMS decarboxylase (ACMSD). As such, ACMSD would be predicted to limit the amount of NAD⁺ produced through *de novo* synthesis. ACMSD is evolutionarily conserved from the nematode worm *C. elegans* to mice⁴ — an observation that is striking because, until recently, nematodes were not thought to synthesize NAD⁺ *de novo*.

The authors inhibited the *acsd-1* gene, which encodes the equivalent of ACMSD in nematodes. This inhibition did increase NAD⁺ levels. Increasing NAD⁺ is well known to extend lifespan in worms, and the authors found that lifespan was longer in the worms in which *acsd-1* expression was completely blocked. Moreover, preventing *acsd-1* expression led to molecular responses that have been linked to defence against ageing^{5,6}: increased activation of the sirtuin enzyme sir-2.1; enhanced mitochondrial function; and a protective mitochondrial stress response.

In mice and humans, ACMSD is most highly expressed in the liver and kidney⁷, and a recent study indicates that these are the main organs for Trp-dependent NAD⁺ generation⁸. Katsyuba *et al.* found that inhibition of the *Acmsd* gene increased NAD⁺

levels and mitochondrial function in cultured mouse liver cells. The authors therefore developed chemical inhibitors of ACMSD, and tested whether these inhibitors could improve outcomes in mouse models of two ageing-related diseases: diet-induced fatty liver disease and acute kidney injury.

Earlier work had already described a beneficial effect of augmenting NAD⁺ in each of these settings^{9,10}. Katsyuba and colleagues' data confirmed the potential for therapeutic NAD⁺ augmentation — treatment with their inhibitors protected against disease in these models. The results also suggest that increases in the *de novo* NAD⁺ synthesis pathway alone are sufficiently robust to ameliorate liver and kidney diseases associated with low NAD⁺ levels. However, proving this will require a demonstration that the benefit of ACMSD inhibition derives from the increase in NAD⁺, rather than from another mechanism such as depletion of the molecule picolinic acid, which is produced by ACMSD-mediated degradation of ACMS. If proved, this finding would be consistent with a study¹¹ that identified a different enzyme in the Trp pathway, quinolinate phosphoribosyltransferase, as a determinant of susceptibility to acute kidney injury.

Several basic questions merit further consideration. For instance, what evolutionary pressures could have led to the conservation of multiple biosynthetic routes to NAD⁺? And why is the *de novo* pathway most active in organs involved in detoxification of the body in mammals? One attractive possibility is that the liver and kidney are more exposed than other organs to toxic stressors that stimulate NAD⁺ consumption. The fact that these organs export Nam to the rest of the body⁸ might explain some aspects of inter-organ metabolic relationships in health and disease — for example, why people with chronic liver disease often develop impaired brain and heart function.

The ACMSD inhibitors developed by Katsyuba *et al.* are indicative of the interest in harnessing NAD⁺ augmentation in the clinic. It has been nearly 20 years since NAD⁺ was first proposed to be a determinant of lifespan¹². But because ageing is so complex, a clinically testable definition has been lacking. Trials to examine the relationship between NAD⁺ augmentation and human lifespan would take too long to be financially feasible. If, instead, a definition of ageing incorporated waning resistance to acute stressors such as infections, trauma or surgery, then clinical testing of NAD⁺ modulators could become more viable. Another study has recently applied this logic, reporting a trial of orally administered Nam among people undergoing cardiac bypass surgery — an invasive procedure often performed on older individuals and associated with post-operative kidney injury¹¹. The beneficial effect of NAD⁺ augmentation on acute kidney injury observed in that work, although preliminary, illuminates a translational track for NAD⁺ manipulation.

However, oral consumption of NAD⁺

precursors might not be an efficient way to increase NAD⁺ levels⁸, so there is a need to consider more-targeted pharmacological approaches. The ACMSD inhibitors developed by Katsyuba and colleagues are therefore a valuable proof of concept. Given the enrichment of enzymes of the *de novo* pathway in the kidney and liver, this particular strategy also raises the intriguing possibility of tissue-specific NAD⁺ manipulation.

The list of conditions potentially amenable to NAD⁺ augmentation is varied and growing, from glaucoma¹³ to neurodegenerative conditions¹⁴ and metabolic syndrome¹⁵. A confluence of work using distinct approaches — human genetics³, radiochemistry⁸, comparative phylogeny¹ and clinical studies¹¹ — now indicates that the Trp pathway is both a major gatekeeper of NAD⁺ levels and a target for medical exploration. ■

Samir M. Parikh is in the Center for Vascular Biology, Department of Medicine and Division

of Nephrology, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts 02215, USA.

e-mail: sparikh1@bidmc.harvard.edu

1. Katsyuba, E. *et al.* *Nature* **563**, 354–359 (2018).
2. Krehl, W. A., Teply, L. J., Sarma, P. S. & Elverhjem, C. A. *Science* **101**, 489–490 (1945).
3. Shi, H. *et al.* *N. Engl. J. Med.* **377**, 544–552 (2017).
4. Fukoka, S.-I. *et al.* *J. Biol. Chem.* **277**, 35162–35167 (2002).
5. Mouchiroud, L. *et al.* *Cell* **154**, 430–441 (2013).
6. Gomes, A. P. *et al.* *Cell* **155**, 1624–1638 (2013).
7. Pucci, L., Perozzi, S., Cimadamore, F., Orsomando, G. & Raffaelli, N. *FEBS J.* **274**, 827–840 (2007).
8. Liu, L. *et al.* *Cell Metab.* **27**, 1067–1080 (2018).
9. Tran, M. T. *et al.* *Nature* **531**, 528–532 (2016).
10. Gariani, K. *et al.* *Hepatology* **63**, 1190–1204 (2016).
11. Poyan Mehr, A. *et al.* *Nature Med.* **24**, 1351–1359 (2018).
12. Lin, S.-J., Defossez, P.-A. & Guarente, L. *Science* **289**, 2126–2128 (2000).
13. Williams, P. A. *et al.* *Science* **355**, 756–760 (2017).
14. Wang, G. *et al.* *Cell* **158**, 1324–1334 (2014).
15. Cantó, C. *et al.* *Nature* **458**, 1056–1060 (2009).

This article was published online on 24 October 2018.

MEDICAL RESEARCH

HIV rebound prevented in monkeys

Antiviral drugs prevent HIV from replicating, but the virus can hide in the cells of infected individuals in a non-replicating, latent form. A two-pronged approach to target this latent virus shows promise in monkeys. SEE ARTICLE P.360

SHARON R. LEWIN

Advances in the management of HIV over the past three decades have been spectacular, thanks to the development of antiretroviral drugs that prevent the virus from replicating. These drugs have very few side effects, prolong life and block sexual transmission. However, the virus is never eliminated — instead, it hides in immune cells called CD4⁺ T cells in a non-replicating, latent form. If treatment is stopped, the virus rapidly re-emerges from this latent reservoir¹. Given the cost of antiretroviral drugs, the need for ongoing engagement in care and the persisting stigma for people living with HIV, there is intense focus on finding a way to target the latent virus so that treatment can be safely stopped without viral re-emergence. On page 360, Borducchi *et al.*² report remarkable findings that may have achieved just that in a monkey model of HIV.

Disappointingly, no intervention has so far managed to eliminate the latent HIV reservoir in people³. Borducchi and colleagues set out to investigate whether a combination of two treatments could do so in monkeys. The first treatment, GS-9620 (vesatolimod), is an oral

drug that activates the Toll-like receptor 7 (TLR7) protein. TLR7, in turn, activates immune cells — not only CD4⁺ T cells, but also CD8⁺ T cells and natural killer (NK) cells, both of which can hunt out and destroy virus-infected cells⁴. Activation of latent HIV contained in CD4⁺ T cells is thought to render them more susceptible to destruction by other immune cells⁵. The second treatment, PGT121, is an antibody, one end of which recognizes and binds to key HIV proteins on the surface of infected cells, with the opposite end triggering other immune cells to destroy the target cell⁶.

Borducchi and colleagues infected 44 monkeys with a hybrid of HIV and the simian immunodeficiency virus. Seven days later, they began to treat the animals with a potent combination of antiretrovirals, similar to that used in humans. HIV rapidly disappeared from the blood of all monkeys, as expected. After 96 weeks, the authors split the monkeys into 4 randomized groups of 11 — one group received no intervention, a second was given GS-9620, a third was injected with PGT121, and a fourth received both GS-9620 and PGT121. The monkeys received these treatments until week 114, then continued to



50 Years Ago

For some years after 1970, West Germany may be expected to have the largest fully steerable radio telescope in the world. A 100 metre (333 feet) instrument is now under construction in the sparsely populated Eifel mountains 40 km south-west of Bonn, and is due to be commissioned in April 1970 ... Work on site clearance and access began about a year ago and if the erection phase goes as smoothly, the Effelsberg telescope will take over leadership as the most powerful precision instrument at short centimetre wavelengths from Jodrell Bank Mark I.

From *Nature* 16 November 1968

100 Years Ago

The turmoil which has shaken the civilised world to its foundations since August, 1914, ceased with the signing of the armistice with Germany on Monday, November 11 ... Let us hope that the immoral militarism which led to the war, and sacrificed all principles of faith-keeping, justice, and humanity to attain its purpose, has been vanquished for ever ... The war has shown that spiritual qualities count for much more than mere numbers. Our system of education was inefficient, but it produced a nation of young heroes ... Though war is not an exact science ... tactics are constantly affected by the progress of science, and disaster may ensue if its effect is not correctly appreciated ... [T]here will be no end to the rich gifts which science will pour into the lap of the human race. Then, if men are worthy of the fruits showered upon them, there will be an end of the night of weeping, and the advent of the morn of song which is our highest heritage. Let us do what we can to hasten the coming of this time, when men shall stretch out their hands to one another and encircle the world.

From *Nature* 14 November 1918

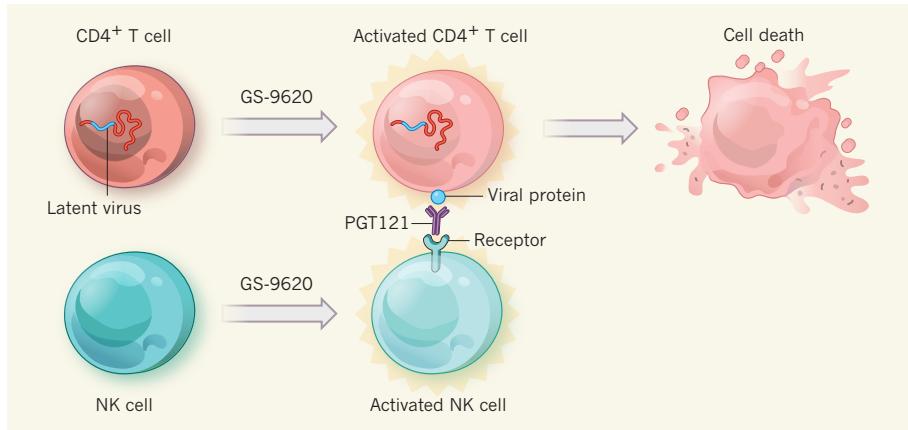


Figure 1 | 'Shock and kill' for latent HIV. Antiretroviral drugs prevent HIV from replicating. However, the virus hides in immune cells called CD4⁺ T cells in a latent, non-replicating form that can re-emerge once antiretroviral treatment is stopped. Borducchi *et al.*² report a two-pronged approach that targets the latent virus during antiretroviral treatment, thus preventing viral rebound. The authors gave monkeys who were infected with a hybrid of HIV and simian immunodeficiency virus and receiving antiretroviral drugs a combination of two treatments — a drug called GS-9620 and an antibody called PGT121. The authors propose that the treatment acts through a mechanism dubbed shock and kill. Under this model, GS-9620 'shocks' CD4⁺ T cells, such that viral proteins become visible on the cell surface. The drug also activates immune cells called natural killer (NK) cells. PGT121 then binds to the viral proteins on the activated infected CD4⁺ T cells. NK cells, in turn, bind to PGT121, and so target infected T cells for destruction.

receive antiretroviral therapy until week 130. The investigators then stopped antiretrovirals and waited to see whether the virus rebounded.

The researchers detected the virus in the blood of all 11 animals that received no intervention, within a median of 21 days after stopping antiretroviral treatment. Viral rebound was also seen in 10 and 9 animals in the groups given only GS-9620 and PGT121, respectively. In stark contrast, only 6 of the 11 monkeys treated with both GS-9620 and PGT121 showed signs of the virus rebounding by week 28 after antiretroviral treatment had ceased. The other 5 monkeys in this group remained completely clear of any detectable virus, even using sensitive assays.

Why was the approach so effective? Borducchi *et al.* found that CD4⁺ T cells and NK cells were activated in all monkeys that received GS-9620. But activating these cells clearly is not sufficient to destroy infected cells, because treatment with GS-9620 alone did not prevent viral rebound, consistent with a previous report⁴. GS-9620 has also been shown to activate latent virus, 'shocking' it out of its hiding place in monkeys to enable targeting by the immune system⁴. The authors did not find evidence for this in the current study, but that might be because they began treating their monkeys soon after infection. This meant that the animals had only a small reservoir of virus, which would be difficult to detect.

Although potent neutralizing antibodies such as PGT121 are being widely tested as a way to prevent HIV infection, it has been unclear whether these antibodies actually kill infected cells in the presence of antiretrovirals. There is an added layer of complexity if the virus is latent — can the antibody even recognize these cells? The authors propose that

GS-9620 treatment activated the CD4⁺ T cells harbouring latent virus, rousing the virus and so allowing the antibody to target the infected cell, perhaps with assistance from activated NK cells that are directed to the cell by the antibody (Fig. 1). This type of approach is referred to as 'shock and kill'.

Borducchi and colleagues' findings are exciting, and offer hope of a cure for HIV, but there are a few reasons for tempered enthusiasm. First, because the authors began treating the monkeys with antiretrovirals extremely soon after infection, the pool of latently infected cells was small and potentially easier to clear than if the virus had had longer to replicate. Most people living with HIV are diagnosed months to years after infection. In the current study, the monkeys that did not rebound were those with the lowest pretreatment viral loads, supporting the idea that a lower burden of virus before treatment might make the reservoir easier to eliminate. Indeed, this has been shown recently in another monkey model⁷.

Second, the hybrid virus used here is potentially easier for the monkey immune system to control than are other monkey viruses⁸. Also, for people infected with HIV, control of virus is extremely rare, even if antiviral treatment is started within days of infection¹. Third, the monkeys were followed for only about six months after antiretroviral treatment was stopped. In people with HIV who stop antiretroviral treatment, rebound of the virus can be delayed for as long as two years^{9,10}, so longer follow-up of these monkeys is needed. Finally, and most importantly, we don't yet know whether interventions in monkey models of HIV reflect what will happen in humans.

The biggest test will now be to see whether administration of GS-9620 and PGT121 (or a

related antibody) can produce similar results in people. These clinical trials are being planned, and the results are eagerly awaited. In the meantime, antiretrovirals remain the best and only option for the long-term treatment of HIV infection. ■

Sharon R. Lewin is at The Peter Doherty Institute for Infection and Immunity,

University of Melbourne and Royal Melbourne Hospital, Melbourne, Victoria 3000, Australia. e-mail: sharon.lewin@unimelb.edu.au

1. Colby, D. J. et al. *Nature Med.* **24**, 923–926 (2018).
2. Borducchi, E. N. et al. *Nature* **563**, 360–364 (2018).
3. Pitman, M. C., Lau, J. S. Y., McMahon, J. H. & Lewin, S. R. *Lancet HIV* **5**, e317–e328 (2018).
4. Lim, S.-Y. et al. *Sci. Transl. Med.* **10**, eaao4521 (2018).
5. Kim, Y., Anderson, J. L. & Lewin, S. R. *Cell Host Microbe* **23**, 14–26 (2018).
6. Bruel, T. et al. *Nature Commun.* **7**, 10844 (2016).
7. Okoye, A. A. et al. *Nature Med.* **24**, 1430–1440 (2018).
8. Nishimura, Y. & Martin, M. A. *Cell Host Microbe* **22**, 207–216 (2017).
9. Henrich, T. J. et al. *PLoS Med.* **14**, e1002417 (2017).
10. Luzuriaga, K. et al. *N. Engl. J. Med.* **372**, 786–788 (2015).

This article was published online on 3 October 2018.

EARTH SCIENCE

Water takes a deep dive into the Mariana Trench

A tectonic plate descending into the Mariana Trench carries sea water deep into Earth's interior. It seems that much more water enters Earth at this location than was thought — with implications for the global water budget. SEE LETTER P.389

DONNA J. SHILLINGTON

The subduction zones at which the tectonic plates beneath the sea thrust into the deep Earth act as gigantic conveyor belts, carrying water, fluids and volatile compounds into our planet. Water in Earth's interior is released back into the oceans and atmosphere by volcanoes. These inputs and outputs constitute a global deep-Earth water cycle, but quantifying the total water input from oceanic plates has proved difficult. On page 389, Cai *et al.*¹ report that the Pacific plate, which subducts in the Mariana Trench, contains much more water than was previously supposed — a finding that has major ramifications for Earth's water budgets.

Water is as crucial to the workings of Earth's interior as it is to Earth's surface processes: among other things, it triggers magma generation beneath volcanoes, lubricates deep fault zones, and fundamentally alters the strength and behaviour of Earth's mantle. Sea water seeps into the oceanic lithosphere through fractures and pores, and reacts with minerals in the crust and mantle to form hydrous minerals (such as serpentine) that store water in their crystal structures.

Water infiltration occurs at a couple of key stages of an oceanic plate's life cycle. The first is at mid-ocean ridges, when water circulates through hot, newly formed oceanic plates². But at fast-spreading ridges (which are the primary 'diet' of the subduction zones that ring the Pacific Ocean), circulation and hydration are mainly restricted to the plate's upper crust. The accumulation of sediments subsequently seals off most of the oceanic plate from the

ocean, but seamounts (underwater mountains) and fracture zones provide pathways for further water input and output, so that circulation continues away from the mid-ocean ridge³. The final infiltration occurs at the 'outer rise' of a subduction zone, where the oceanic plate bends before entering the trench. Here, extensional faults form in response to bending, and are thought to enable pervasive, deeply penetrating hydration of the crust and upper mantle^{4–6}.

The evidence for water entering subduction zones is clear, but several knowledge gaps have hindered attempts to quantify the total volume of water going down these hatches, even at individual subduction zones. One unknown is the depth to which water penetrates the

oceanic plate. Most constraints on estimates come from controlled-source seismic data, which are produced by measuring seismic waves generated by artificial sources using dense arrays of recording instruments. These data provide excellent constraints on hydration of the crust and shallow mantle, but, with one notable exception⁷, do not constrain the full depth of hydration. It is clearly not possible to tally the total volume of subducted water without knowing the full hydration depth.

Another important challenge is to untangle all the factors that alter the speed at which seismic waves travel through different parts of the plate; measurements of such seismic waves are one of the primary means of estimating the amount of water in the subducting oceanic plate. Most estimates assume that any reduction in wave speed results from the replacement of olivine (the main mineral found in the mantle) by serpentine. However, in the crust and shallow mantle, water-filled cracks can also contribute to velocity reductions⁸. To complicate things further, seismic-wave speeds in the upper oceanic mantle are anisotropic — they depend on the direction of propagation. This is because olivine crystals align in the direction in which the sea floor spreads when new oceanic plates are created at mid-ocean ridges. Further anisotropy can result from fractures formed at the outer rise.

Cai *et al.* tackle all of these issues by presenting constraints on the hydration of the approximately 150-million-year-old Pacific plate as it subducts at the Mariana Trench. The authors analysed seismic waves from distant earthquakes, recorded by an array of seismometers on the sea floor. This allowed them to model seismic-wave speeds to much greater depths (albeit at lower resolution) than is possible using controlled-source seismic data.

The researchers find that, impressively, the full hydration depth of the lithosphere extends to approximately 30 kilometres below the sea floor (Fig. 1). They were also able to examine velocity reductions in deep regions at which the pressure would be sufficiently high to close all cracks, thus allowing them to eliminate the possible contribution of such cracks to velocities in these regions. Finally, because the authors recorded waves travelling in all directions across their array, they were able

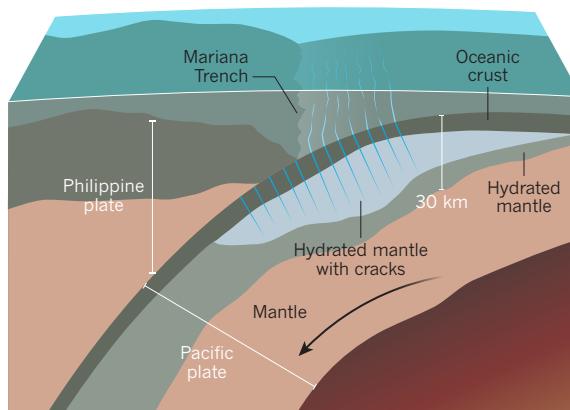


Figure 1 | Hydration of the Pacific tectonic plate at the Mariana subduction zone. At the Mariana Trench in the Pacific Ocean, the Pacific plate slips (subducts) beneath the adjacent Philippine plate, transporting sea water into the deep Earth. The water seeps through cracks and pores in the plate, and reacts with minerals in the crust and mantle to form hydrated regions consisting of minerals that store water in their crystal structures. Cai *et al.*¹ have used seismic measurements to show that water penetrates to depths of about 30 kilometres below the ocean floor. (Approximated from Fig. 2d of ref. 1.)

to account for the contribution of anisotropy to wave speed.

Cai *et al.* report that more than four times as much water is entering the Mariana subduction zone than was previously estimated⁹. Old, cold subducting plates such as that entering the Mariana Trench are particularly effective conveyors of water into the deep Earth because hydrous minerals in such plates are stable to greater depths than in younger, hotter plates. If extrapolated globally to other places where old, cold plates subduct, the authors' result implies that the amount of water entering Earth's interior greatly exceeds current estimates¹⁰ of the amount being emitted by volcanoes, and thus requires a rethink of the global water budget.

Several questions still need to be answered to better constrain estimates of the inputs to Earth's deep-water cycle and evaluate the implications of the new results. First, how variable is the water content in the oceanic plate at a range of depths and scale lengths along subduction zones? Many studies have reported changes in extensional faulting and in crustal and upper-mantle hydration along subduction zones, and several competing factors have been proposed to contribute to these changes^{7,11,12}. Characterizing this variability throughout the hydrated part of the plate and understanding its causes will be essential for us to tally water inputs and compare them with outputs. It would also be useful to determine whether hydration occurs in focused zones near faults, as has been observed in a different tectonic setting¹³, or is distributed more evenly, because this might control whether mineral-bound water is released by dehydration or is carried to greater depths¹⁴.

Finally, because a substantial volume of water is probably stored in the crust and upper mantle (the regions that are most accessible to seawater infiltration), the thorny issue of whether changes to seismic waves reflect the presence of water-filled cracks or of hydrous minerals still needs to be directly addressed in these areas. A comprehensive understanding of inputs to Earth's interior will require a multi-pronged approach, including multi-scale marine geophysical studies, drilling of the faults that are thought to be conduits for water into the oceanic lithosphere, and numerical and laboratory studies. But for now, Cai and colleagues' results have taken us a big step closer to understanding the total water input at subduction zones. ■

Donna J. Shillington is at the Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York 10964, USA.
e-mail: djs@ldeo.columbia.edu

1. Cai, C., Wiens, D. A., Shen, W. & Eimer, M. *Nature* **563**, 389–392 (2018).
2. Alt, J. C. *Geophys. Monogr.* **91**, 85–114 (1995).
3. Fisher, A. T. *et al.* *Nature* **421**, 618–621 (2003).
4. Ranero, C. R., Phipps Morgan, J., McIntosh, K. & Reichert, C. *Nature* **425**, 367–373 (2003).

5. Ivandic, M., Grevermeyer, I., Berhorst, A., Flueh, E. R. & McIntosh, K. *J. Geophys. Res.* **113**, B05410 (2008).
6. Lefeldt, M., Grevermeyer, I., Goßler, J. & Bialas, J. *Geophys. J. Int.* **178**, 742–752 (2009).
7. Van Avendonk, H. J. A., Holbrook, W. S., Lizarralde, D. & Denyer, P. *Geochem. Geophys. Geosys.* **12**, Q06009 (2011).
8. Miller, N. C. & Lizarralde, D. *Geophys. Res. Lett.* **43**, 7982–7990 (2016).
9. van Keken, P. E., Hacker, B. R., Syracuse, E. M. & Abers, G. A. *J. Geophys. Res.* **116**, B01401 (2011).
10. Parai, R. & Mukhopadhyay, S. *Earth Planet. Sci. Lett.* **317–318**, 396–406 (2012).
11. Shillington, D. J. *et al.* *Nature Geosci.* **8**, 961–964 (2015).
12. Fujie, G. *et al.* *Nature Commun.* **9**, 3844 (2018).
13. Bayrakci, G. *et al.* *Nature Geosci.* **9**, 384–388 (2016).
14. Wada, I., Behn, M. D. & Shaw, A. M. *Earth Planet. Sci. Lett.* **353–354**, 60–71 (2012).

ORGANIC CHEMISTRY

From hydrocarbons to precious chemicals

Operationally simple chemistry enables aliphatic carbon–carbon bonds – the ‘girders’ in the framework of many organic molecules – to be prepared from widely available hydrocarbons known as alkenes. SEE LETTER P.379

JAMES P. MORKEN

Many valued reagents and catalysts used in the preparation of organic compounds are highly reactive and are therefore incompatible with exposure to the open atmosphere – in some cases, dangerously so. The instrumentation required to carry out reactions with these compounds, such as high-vacuum apparatus and gloveboxes (isolation chambers), is costly and demands special training. Chemical processes that do not have such requirements are therefore more likely to make a big impact on how chemists synthesize molecules, be it for the development of new

materials, pharmaceuticals or agrochemicals. The reactions reported by Fu and co-workers¹ on page 379 are a case in point. Not only are they simple to carry out, but they also deliver a variety of useful products that are otherwise much more difficult to make.

Parallels are often drawn between the fields of organic synthesis and architecture²: aliphatic carbon–carbon (C–C) bonds are the architect's ‘girders’ on which many structurally complex molecules are built. Installing these girders is challenging, and necessitates the use of highly reactive reagents. To add to the challenge, the orientation in which new C–C bonds are installed – the stereochemistry of

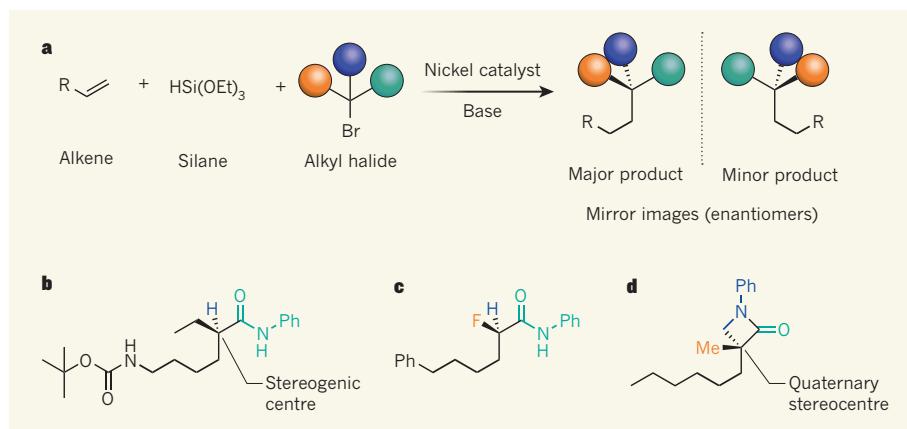


Figure 1 | Operationally simple reactions for making aliphatic carbon–carbon bonds. **a**, Fu and colleagues¹ report chemistry in which an alkene, a silane and an alkyl halide react in the presence of a nickel catalyst and a base to form potentially useful products. The reactions can be carried out without excluding air or moisture, which makes them straightforward in practice. Moreover, the reactions are enantioselective: they produce one isomer of the reaction product to the near exclusion of the product's mirror-image isomer. R and the coloured spheres represent a variety of organic groups or atoms; Si, silicon; Et, ethyl group; Br, bromine. **b–d**, The authors prepared a variety of products, including **b**, compounds in which a stereogenic centre (a carbon atom that has three other groups attached by carbon atoms) is next to a carbonyl group (C=O); **c**, compounds that contain fluorine atoms; and **d**, compounds that contain quaternary stereocentres (carbon atoms attached to four different groups by carbon atoms). Ph, phenyl group; Me, methyl group.

the reaction — affects the overall shape of the final molecule³, which in turn can affect the molecule's function in applications.

Fu and colleagues' advance addresses these challenges. The authors describe a new C–C bond-forming reaction, known as a cross-coupling reaction, that produces one isomer of the reaction product to the near exclusion of the product's mirror-image isomer (in chemists' terms, the reaction is said to be enantioselective). Moreover, the process does not require the use of highly reactive and fragile reagents.

The authors' approach requires three reagents: an alkene, a silane and an alkyl halide (Fig. 1a). Alkenes are not sensitive to air, which distinguishes Fu and colleagues' reactions from the majority of cross-coupling reactions⁴, in which the alkene is replaced by an air-sensitive organometallic compound, either as a reagent or as the precursor to a reagent. The new reactions seem to involve an orchestrated set of events wherein the alkene first attaches to a catalytic nickel complex, which is generated *in situ* by a process that involves the silane reagent. The attachment of the alkene produces a transient reactive species, which then reacts with the alkyl halide to form the new C–C bond.

Fu and co-workers' nickel-catalysed process is related to one reported⁵ by another team in 2016, but enhances the usefulness of that approach by addressing two key challenges. First, a catalyst had to be identified that not only promotes stereoselective C–C bond formation for an array of different substrates, but also activates the alkene without promoting side reactions between the silane and either the alkyl halide or the alkene. Second, reaction conditions had to be identified that allowed a base to drive catalytic cycles — which is difficult in this context, because bases often interconvert mirror-image isomers.

Not only is the use of alkenes as replacements for reactive organometallic reagents appealing in terms of its practical simplicity, but it also broadens the range of substrates that can be used in Fu and colleagues' reactions. Alkenes are widely available, many are produced industrially on a large scale, and they can be generated by a variety of chemical processes. Alkenes are also especially attractive as reagents for chemical synthesis: they are chemically inert to a range of reagents, but can be induced to react in the presence of the right catalyst and under the right set of conditions. Impressively, the catalytic conditions used by the authors allow alkenes to react without interference when a variety of other common organic groups are also attached to the alkene (see Fig. 2a of the paper¹).

Intriguingly, when Fu and colleagues used internal alkenes — in which the characteristic carbon–carbon double bond of the alkene is in the middle of a chain of carbon atoms — in their reactions, they observed a phenomenon called chain-walking⁶, which causes the double

bond to migrate to the end of the carbon chain before reacting. This observation means that products obtained from an increasingly used type of reaction known as olefin cross-metathesis⁷ (which produces internal alkenes) might be suitable substrates. It will also be exciting to find out whether the step in which the alkene attaches to the nickel complex can be made to be enantioselective, because this would allow products containing multiple stereogenic centres (carbon atoms to which three different groups are attached by carbon atoms) to be generated enantioselectively.

A particularly notable feature of Fu and co-workers' strategy is that a considerable array of alkyl halides can be used, some of which are not effective substrates for cross-coupling reactions with organometallic reagents. For example, the authors use alkyl halides known as secondary α -halo amides in their reactions, and show that this provides a simple and enantioselective route to prepare compounds that contain a carbonyl (C=O) group next to a stereogenic centre (Fig. 1b). Such compounds are potentially versatile intermediates for chemical synthesis, and have most commonly been prepared using a much less efficient approach based on the use of compounds called chiral auxiliaries⁸. The researchers also demonstrate that they can use their enantioselective reactions to make certain fluorine-containing compounds (see Fig. 1c, for example), which might be useful in

medicinal chemistry. Moreover, the chemistry can be used to make compounds that contain quaternary stereocentres (Fig. 1d) — carbon atoms to which four different groups are attached by carbon atoms, which are some of the most difficult structures to prepare enantioselectively.

Overall, this advance is a much-needed method for the enantioselective synthesis of an impressive assortment of versatile small organic molecules, many of which will be of value to research at the frontiers of chemistry. ■

James P. Morken is in the Department of Chemistry, Boston College, Chestnut Hill, Massachusetts 02467, USA.
e-mail: james.morken@bc.edu

1. Wang, Z., Yin, H. & Fu, G. C. *Nature* **563**, 379–383 (2018).
2. Trauner, D. *Angew. Chem. Int. Edn* **57**, 4177–4191 (2018).
3. Nguyen, L. A., He, H. & Pham-Huy, C. *Int. J. Biomed. Sci.* **2**, 85–100 (2006).
4. Magano, J. & Dunetz, J. R. *Chem. Rev.* **111**, 2177–2250 (2011).
5. Lu, X. *et al.* *Nature Commun.* **7**, 11129–11136 (2016).
6. Vasseur, A., Bruffaerts, J. & Marek, I. *Nature Chem.* **8**, 209–216 (2016).
7. Hoveyda, A. H. & Zhugralin, A. R. *Nature* **450**, 243–251 (2007).
8. Kohler, M. C., Wengryniuk, S. E. & Coltart, D. M. in *Stereoselective Synthesis of Drugs and Natural Products* (eds Andrushko, V. & Andrushko, N.) 183–213 (Wiley, 2013).

HUMAN DEVELOPMENT

The landscape of early pregnancy

RNA sequencing of thousands of single cells located at the interface between mother and fetus in early pregnancy reveals remarkable complexity in the cell types and regulatory networks that support reproduction. SEE ARTICLE P.347

SUMATI RAJAGOPALAN & ERIC O. LONG

Scientists have long puzzled over the 'immunological paradox' of pregnancy¹: how does the mother tolerate the fetus — a foreign entity that carries some of the father's DNA? On page 347, Vento-Tormo *et al.*² investigate this enigma. The authors performed single-cell RNA sequencing (scRNAseq) of cells isolated from the placenta and the decidua (the lining of the pregnant uterus), and from matching maternal blood for comparison. They identified an array of cell types unique to this maternal–fetal interface, and inferred the existence of a large network of potential interactions between them that would favour immunological tolerance and nurture the

growth of the fetus. The authors' molecular atlas provides an impressive resource for future studies of pregnancy and its complications.

The early embryo develops into a structure called the blastocyst, which implants in the lining of the uterus. Implantation triggers the development of the placenta from fetal membranes. The placenta nourishes the fetus through the umbilical cord³. Abnormal placental development can lead to several complications of pregnancy, including pre-eclampsia, fetal growth restriction and stillbirth. A better understanding of human placental development is sorely needed, but there is no good animal model for this process — it has to be studied in women.

Vento-Tormo *et al.* collected placental,

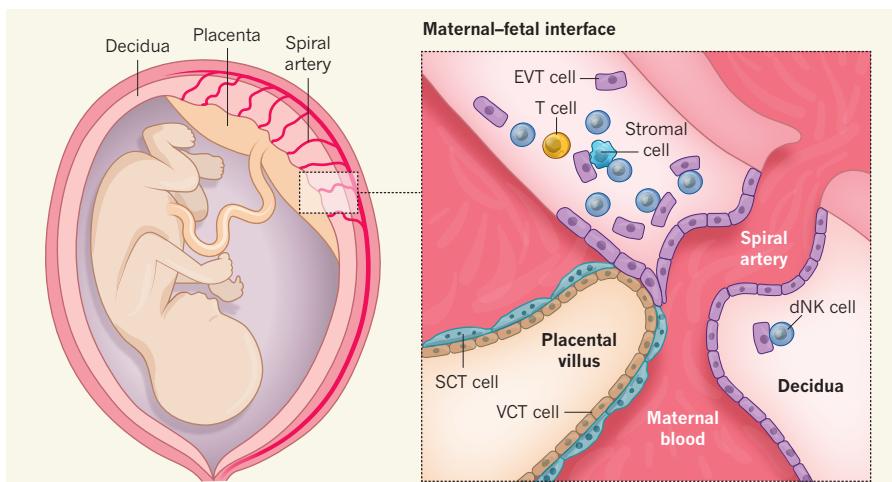


Figure 1 | An atlas of cells at the maternal–fetal interface. During the first trimester of human pregnancy, an interface forms between the maternal decidua (the lining of the pregnant uterus) and the fetal placenta. Nutrients are delivered to the placenta down maternal spiral arteries. Vento-Tormo *et al.*² sequenced the RNA of thousands of single cells at this interface, and used the data to define different cell types and to predict interactions between cells on the basis of the receptors and ligand molecules that they express (not shown). The authors' data provide information about fetal cell types derived from the early embryo: villous cytotrophoblast (VCT) cells, which line placental structures called villi; syncytiotrophoblast (SCT) cells that cover the villus surface; and extravillous trophoblast (EVT) cells, which line the maternal blood vessels and intermingle with maternal cells in the decidua. The authors also identified several types of maternal immune cell, including T cells and three subsets of decidual natural killer (dNK) cell, and three types of stromal cell, which provide structural support for the decidua.

decidual and blood samples from pregnancies that had been electively terminated at between 6 and 14 weeks of gestation. The authors' scRNAseq analysis enabled them to distinguish between cells of maternal and fetal origin, because the latter include RNA sequences that are absent in the mother. This clearly revealed that cells from the fetus had migrated into the maternal decidua (Fig. 1), and that a small subset of maternal immune cells called macrophages were located in the placenta.

The blastocyst-stage embryo takes an active role in its own destiny. Cells from the outer layer of the blastocyst, called trophoblast cells, undergo differentiation. Vento-Tormo and colleagues identified the transcription factors involved in the differentiation of one type of trophoblast cell, villous cytotrophoblast (VCT) cells, into either syncytiotrophoblast cells or extravillous trophoblast (EVT) cells (Fig. 1). The authors found that VCT cells express receptors that promote differentiation and are stimulated by growth factors produced by various placental cells. EVT cells invade the decidua, where they interact with maternal white blood cells to trigger the remodelling of narrow maternal spiral arteries into wider conduits that can meet the nutritional needs of the developing fetus. The authors showed that such invading EVT cells produce a signalling protein called transforming growth factor β , which favours the development of maternal regulatory T cells — a subset of immune cells called T cells — that rein in immune responses.

The most abundant maternal immune cells in the decidua during the first trimester of pregnancy are natural killer (NK) cells⁴. Best known

as killers of infected cells and tumour cells, NK cells assume a more peaceful role in pregnancy, secreting soluble proteins that promote maternal blood-vessel remodelling^{3,5}. Decidual NK (dNK) cells also regulate the extent to which EVT cells can invade the decidua⁴. Vento-Tormo *et al.* identified three subsets of dNK cell — a remarkable finding, because it shows that dNK cells have evolved into specialized cells that are very different from blood NK cells. The authors' data indicate that the immunological activity of each dNK subset is dictated by their ability to interact with both maternal and fetal cells in the decidua, with the dual outcome of promoting fetal growth and restraining immune attack on the fetal cells.

The researchers' work also reveals that the decidua's two layers are defined by distinct molecular profiles, and contain different complements of five cell types: two types of perivascular cell, which support the maternal blood vessels, and three types of decidual stromal (dS) cell, which provide tissues with structural support. The dS cells express the protein interleukin-15, which is essential for NK-cell survival and proliferation, and ligands for two inhibitory receptors on NK cells, indicating the role of dS cells in supporting the survival of NK cells while restraining their immune function.

To analyse their very large data sets, Vento-Tormo *et al.* created a computational platform, CellPhoneDB, to statistically predict receptor-ligand pairs between the different cell types identified by scRNAseq. The platform is publicly available (CellPhoneDB.org) as a resource for examining gene-expression profiles of single cells and for making inferences

about networks of cell–cell communication. The authors have highlighted just a few of the cellular interactions revealed by their analysis. Many more remain, and await interrogation.

There are inherent limitations to the study of human reproduction. In this case, samples from pregnancies at 6–14 weeks' gestation were treated as equivalent. But during this time, the fetus is nourished in two distinct ways — first, by glands in the uterus that feed into the intravillous space of the placenta, and, later, by the maternal blood, which passes directly to the developing placenta⁶. Treating these two phases as one might obscure valuable information. Furthermore, changes that occur during earlier stages of embryo development were not examined. Obviously, systematic longitudinal analysis of *in utero* human development is not feasible, because of ethical issues.

A major limitation to understanding human development has been the lack of representative animal models. Vento-Tormo *et al.* now provide a human molecular reference against which pregnancy in animals can be analysed to find features that are shared with humans. In addition, data obtained from women with complications of pregnancy can be assessed using this resource. This could lead to the identification of biomarkers of common pregnancy complications.

By mapping the cellular and molecular terrain of the first trimester of human pregnancy, the current study illuminates how the maternal–fetal interface is a peaceful and tolerant environment in which immunological reactivity is dampened. In such a milieu, maternal and fetal cells cooperate to regulate trophoblast invasion, remodel the maternal vasculature and provide sufficient nourishment for the fetus. However, this immunological tolerance might come at a cost. For instance, the well-known vulnerability to certain infections⁷, such as cytomegalovirus, Zika virus and malaria-causing parasites, during this time in pregnancy could be due to restrained immune reactivity. Vento-Tormo and colleagues' data provide a powerful framework in which to assess the landscape of early pregnancy during such devastating infections. ■

Sumati Rajagopalan and Eric O. Long are at the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, Maryland 20892-9418, USA.
e-mails: sumi@nih.gov; elong@niaid.nih.gov

1. Medawar, P. B. *Symp. Soc. Exp. Biol.* **7**, 320–338 (1953).
2. Vento-Tormo, R. *et al.* *Nature* **563**, 347–353 (2018).
3. Erlebacher, A. *Annu. Rev. Immunol.* **31**, 387–411 (2013).
4. Moffett, A. & Colucci, F. *J. Clin. Invest.* **124**, 1872–1879 (2014).
5. Rajagopalan, S. *Cell. Mol. Immunol.* **11**, 460–466 (2014).
6. Burton, G. J., Watson, A. L., Hempstock, J., Skepper, J. N. & Jauhia, E. *J. Clin. Endocrinol. Metab.* **87**, 2954–2959 (2002).
7. Yockey, L. J. & Iwasaki, A. *Immunity* **49**, 397–412 (2018).

Anthropogenic influences on major tropical cyclone events

Christina M. Patricola^{1*} & Michael F. Wehner²

There is no consensus on whether climate change has yet affected the statistics of tropical cyclones, owing to their large natural variability and the limited period of consistent observations. In addition, projections of future tropical cyclone activity are uncertain, because they often rely on coarse-resolution climate models that parameterize convection and hence have difficulty in directly representing tropical cyclones. Here we used convection-permitting regional climate model simulations to investigate whether and how recent destructive tropical cyclones would change if these events had occurred in pre-industrial and in future climates. We found that, relative to pre-industrial conditions, climate change so far has enhanced the average and extreme rainfall of hurricanes Katrina, Irma and Maria, but did not change tropical cyclone wind-speed intensity. In addition, future anthropogenic warming would robustly increase the wind speed and rainfall of 11 of 13 intense tropical cyclones (of 15 events sampled globally). Additional regional climate model simulations suggest that convective parameterization introduces minimal uncertainty into the sign of projected changes in tropical cyclone intensity and rainfall, which allows us to have confidence in projections from global models with parameterized convection and resolution fine enough to include tropical cyclones.

Tropical cyclones are among the deadliest and most destructive natural disasters. Hurricane Katrina was the costliest natural disaster in the USA, causing at least 1,833 deaths and costing US\$160 billion in damages (all dollars adjusted to 2017) along the Gulf Coast of the USA in August 2005¹. A close second is hurricane Harvey, which stalled over the Houston metropolitan area in August 2017, causing record flooding. Hurricane Harvey was followed in September by hurricane Irma, which heavily affected the Virgin Islands and Florida Keys, and hurricane Maria, which caused lasting devastation in Puerto Rico. In total, the hyperactive 2017 Atlantic hurricane season caused at least US\$265 billion in damages and 251 fatalities, probably a staggering underestimate owing to crippled communications and infrastructure in Puerto Rico, which meant that many hurricane-related deaths were unconfirmed. To improve the resiliency of coastal and island communities, it is critical to understand the drivers of tropical cyclone variability and change. However, the response of tropical cyclone activity to climate change, so far and in the future, remains uncertain^{2–4}.

There is no consensus regarding whether climate change until now has influenced tropical cyclone activity, given that natural variability is large and tropical cyclone observation methodologies have changed over time. As yet, there has been no detectable trend in tropical cyclone frequency. Although a positive trend in Atlantic tropical cyclone number has been observed since 1900, it is due primarily to increases in short-lived tropical cyclones, which were probably undersampled during the pre-satellite era when observations over ocean were taken by ship⁵. Subjective measurements and variable observation procedures pose serious challenges to detecting trends in tropical cyclones⁶. This is apparent even when observations are adjusted in attempt to normalize for changing sampling procedures over time, owing to large natural variability⁷. In addition, it remains inconclusive whether there have yet been trends in global tropical cyclone intensity, with increases detected in some basins^{8–11}. The strong influences on tropical cyclones of multi-decadal variability including the Atlantic Multidecadal Oscillation^{12,13} and interannual variability including the

El Niño–Southern Oscillation and Atlantic Meridional Mode^{14–17} make disentangling the influences of climate variability and change on trends in tropical cyclone activity all the more challenging.

Looking into the future, there is no consensus regarding how anthropogenic emissions are expected to change global tropical cyclone frequency, with the majority of climate models projecting fewer tropical cyclones^{18–22} but others more tropical cyclones²³ (see also references within refs^{2–4}). However, maximum potential intensity theory and recent climate modelling studies suggest increases in the future number of intense tropical cyclones^{21,22,24–29}. In addition, climate model simulations suggest that rainfall associated with tropical cyclones will increase in future warmer climates, but with large uncertainty in magnitude^{18,20,28,30–33}. The Clausius–Clapeyron relation dictates that the saturation specific humidity of the atmosphere increases by 7% per 1 °C of warming, providing a constraint on changes in moisture available for precipitation. If tropical cyclone precipitation efficiency does not change, then changes in precipitation follow Clausius–Clapeyron scaling as the oceans warm³⁴. However, recent studies of hurricane Harvey found 15%–38% increases in storm total precipitation attributable to global warming, well above the Clausius–Clapeyron limit of 7% given anthropogenic warming of 1 °C in the Gulf of Mexico^{35–37}. Such rainfall over Houston—a once-every-2,000-year event in the late twentieth century—is expected to become a more common one-per-century event by the end of the twenty-first century³⁸.

There is no theory of tropical cyclone formation to predict how tropical cyclones are expected to change in the future, and the problem is complicated by potentially compensating influences of greenhouse gases. Although the factors that influence tropical cyclones are well understood, with favourable conditions including a warm upper-ocean temperature, an unstable atmosphere with a moist mid-troposphere, and weak vertical wind shear³⁹, the way in which these factors will change, and which ones will dominate, is unknown. Sea-surface temperature (SST) warming has been observed and is expected to continue, which would intensify tropical cyclones⁸. However, sub-surface ocean

¹Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ²Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. *e-mail: cmpatricola@lbl.gov

Table 1 | Tropical cyclone peak 10-m wind speed

Basin	Tropical cyclone	Resolution	Historical minus pre-industrial	RCP4.5 minus historical	RCP6.0 minus historical	RCP8.5 minus historical	Historical	Observed
Atlantic	Katrina	27 km (P)	−1.0			11.0**	101	150
		9 km (P)	2.0			15.2**	123	150
		9 km	−0.5			13.5**	127	150
		3 km	−2.4			13.7**	149	150
		4.5 km		6.0**	8.5**	13.8**	142	150
	Irma	4.5 km	−1.9	7.3**	10.4**	12.4**	143	160
	Maria	4.5 km	−1.5	7.5**	10.9**	12.9**	132	150
	Andrew	4.5 km		−3.3	−2.4	−1.7	118	150
	Bob	4.5 km		−6.1**	−2.4*	2.1	78	100
	Floyd	4.5 km		11.2**	13.5**	X	118	135
Eastern Pacific	Gilbert	4.5 km		18.0**	18.6**	28.8**	109	160
	Ike	4.5 km		12.8**	14.1**	18.0**	127	125
	Matthew	4.5 km		10.6**	11.1**	15.8**	123	145
	Iniki	4.5 km		−0.4	−3.9	4.6*	114	125
	Northwest Pacific	Haiyan	4.5 km	6.7**	3.8	12.3**	124	170
		Morakot	4.5 km	0.5	X	X	71	80
		Songda	4.5 km	10.4**	5.5**	X	109	125
South Pacific	Yasi	4.5 km		11.2**	13.7**	18.9**	95	135
	Southwest Indian	Gafilo	4.5 km	8.6**	8.8**	16.8**	110	140

The ensemble-mean difference in tropical cyclone peak 10-m wind speed is given (in knots) between the historical and pre-industrial simulations and between the RCP4.5, RCP6.0 and RCP8.5 simulations and the historical simulation, along with the tropical cyclone peak 10-m wind speed from observations and the ensemble-mean historical simulation. Cases of substantial differences between simulated and observed tropical cyclone tracks are denoted X (see Methods) and simulations that were not performed are blank. *Changes significant at the 10% level; **changes significant at the 5% level. Simulations that used convective parameterization are denoted 'P'.

structure changes are also important for tropical cyclone intensity, and may be a dampening effect in the future⁴⁰. Considering atmospheric factors, anthropogenic warming is expected to be greater in the upper compared to lower troposphere in response to increased greenhouse gases, which could weaken tropical cyclones. However, the tropical tropopause is expected to cool as its height increases, which would strengthen the maximum potential intensity of tropical cyclones, as observed in the Atlantic^{41,42}. Because maximum potential intensity theory applies to mature tropical cyclones, this means the strength of intense tropical cyclones may increase. In addition to thermodynamic influences on tropical cyclones, changes in atmospheric circulation are also important. Projected increases in vertical wind shear could work to suppress tropical cyclones regionally⁴³. Finally, it is uncertain how the seedling disturbances that serve as tropical cyclone precursors may change.

Observational consistency issues and compensating physical mechanisms for tropical cyclone changes are only some of the challenges in understanding anthropogenic influences on tropical cyclones. In addition, it can be difficult for climate models to represent the observed climatology of intense tropical cyclones, even at the 0.25° horizontal resolution that is considered to be high resolution for global models²¹. Furthermore, the decades-long simulations used to project future tropical cyclone activity typically parameterize convection. However, the associated uncertainty introduced into tropical cyclone projections has not been systematically understood.

The purpose of this study is to advance our understanding of anthropogenic influences on tropical cyclones by quantifying the impact of climate change so far, and in the future, on the intensity and rainfall of destructive tropical cyclone events using convection-permitting regional climate model simulations. We first addressed the question of how tropical cyclone intensity and rainfall could change if hurricanes like Katrina, Irma and Maria occurred in pre-industrial or future warmer climates. We then investigated the robustness of our results by extending the analysis to 15 tropical cyclone events sampled globally under three future climates. Finally, we quantified the uncertainty in these estimates associated with convective parameterization for hurricane Katrina.

Convection-permitting tropical cyclone simulations

We performed simulations with the Weather Research and Forecasting (WRF) regional climate model, which is developed by the National Center for Atmospheric Research (NCAR). Control simulations for each tropical cyclone event consist of ten-member ensemble hindcasts representing the historical conditions in which the tropical cyclone actually occurred, with boundary conditions from reanalysis or observations (Methods). We also performed experiments representing hurricanes Katrina, Irma and Maria if they were to occur in a pre-industrial climate, as well as 15 tropical cyclone events sampled globally at the end of the twenty-first century under the Representative Concentration Pathways emissions scenarios RCP4.5, RCP6.0 and RCP8.5 (listed in Table 1). Although previous studies considered individual tropical cyclones, the modelling frameworks used differ (Methods). Our use of one model for many events allows us to assess the robustness of the climate change responses more directly among events. We selected tropical cyclones that were particularly destructive (in terms of fatalities and economic losses) and represent various tropical cyclone basins. Many of the tropical cyclones were intense in terms of wind speed, but two were weak tropical cyclones with moderate-to-heavy rainfall (typhoon Morakot and hurricane Bob). Boundary conditions for the pre-industrial and Representative Concentration Pathway experiments were based on those from the historical, adjusted to remove and add, respectively, the thermodynamic component of climate change (Methods). Simulations of all tropical cyclones were performed at a convection-permitting horizontal resolution of 4.5 km. To investigate uncertainty in the response of tropical cyclones to anthropogenic forcings caused by convective parameterization, we performed additional simulations of hurricane Katrina at horizontal resolutions of 3 km, 9 km (both without and with parameterization) and 27 km.

Anthropogenic influences on tropical cyclone intensity

To evaluate anthropogenic influences on hurricanes Katrina, Irma and Maria, we first verified that the hindcasted tropical cyclone tracks reasonably represent the observed tracks. Indeed, this is the case for each ensemble member of the historical simulations, as well as the ensemble

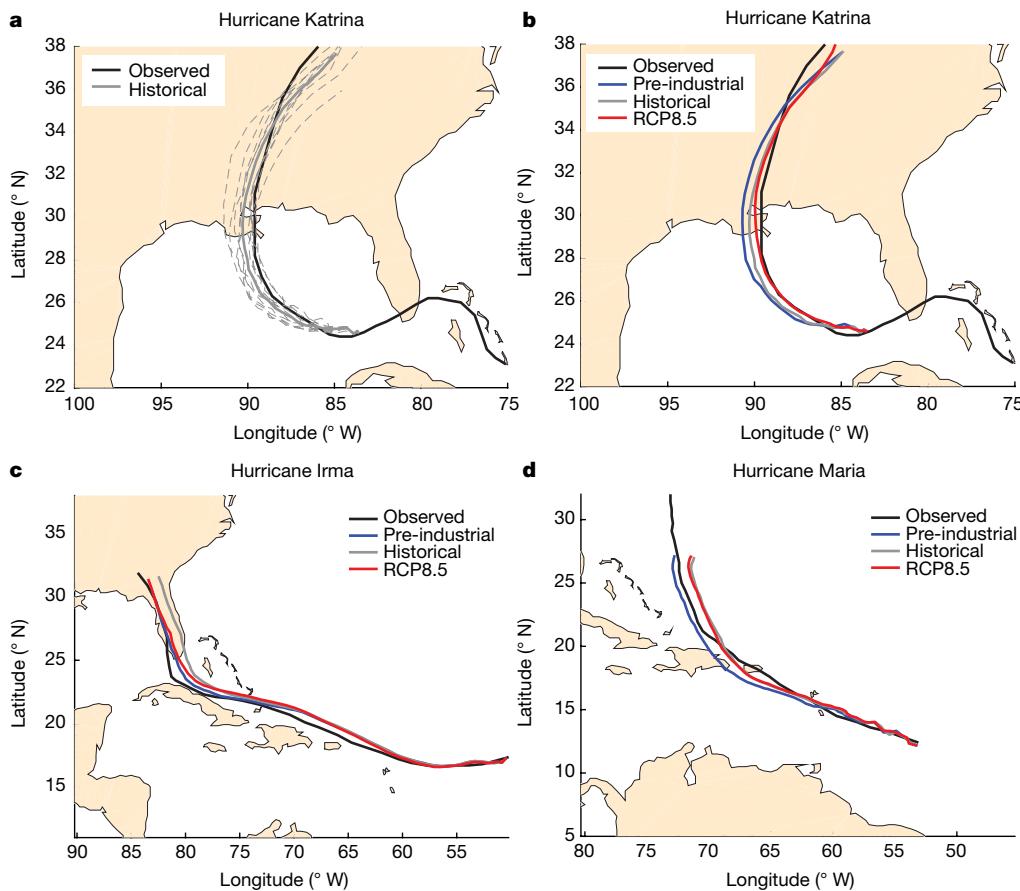


Fig. 1 | Tropical cyclone tracks. **a, b**, Hurricane Katrina's observed track (black) with simulated tropical cyclone tracks from ten ensemble members (grey dashed line) and the ensemble mean (grey solid line) of the historical

simulation (a) and the ensemble mean of historical (grey), pre-industrial (blue) and RCP8.5 (red) simulations at 3-km resolution (b). **c, d**, As in b, for hurricanes Irma (c) and Maria (d) at 4.5-km resolution.

means (Fig. 1 and Extended Data Fig. 1). (We note a bias in hurricane Irma's landfall that is more noticeable owing to Florida's longitudinally narrow geography; also the simulated hurricane Maria slightly missed direct landfall over Puerto Rico.) In addition, the simulated tropical cyclone tracks are robust to anthropogenic perturbations (Fig. 1b–d), indicating that comparisons among experiments are fair (Methods). Next, we estimated the model's ability to simulate the observed intensity of the tropical cyclones, recognizing the challenges of such observation–model comparison (Methods). The time series of maximum 10-m wind speed (Fig. 2a–c) and minimum sea-level pressure (SLP) (Extended Data Fig. 2a–c) show that the hindcasted intensity is close to what was observed for hurricane Katrina, but underestimated for hurricanes Maria and Irma. In addition, a period of rapid intensification was observed for all three hurricanes, which was most pronounced for hurricane Maria. However, the hindcasts failed to represent rapid intensification, a challenge that remains in operational forecasting⁴⁴.

Given that the simulated hurricane tracks and intensities compare reasonably well with the observed tracks, albeit with a failure to reproduce rapid intensification, we evaluated the response in hurricane intensity to past and future anthropogenic forcings. For each of the hurricanes, the ensemble-mean wind speed and SLP-based intensity time series are indistinguishable between the pre-industrial and historical simulations, whereas there is a distinct increase in intensity from the historical to RCP8.5 climates for a substantial portion of each hurricane's lifetime (Fig. 2a–c and Extended Data Fig. 2a–c). To assess the significance ($P = 0.05$) of the intensity changes, we calculated the peak intensity over each hurricane's lifetime based on maximum wind speed (Fig. 2d–f) and minimum SLP (Extended Data Fig. 2d–f) for each ensemble member of the pre-industrial, historical and RCP8.5 simulations. We found that climate change at the time of the event weakly and insignificantly ($P = 0.05$) influenced the intensity of

hurricanes Katrina, Irma and Maria (Table 1 and Extended Data Fig. 3), corresponding to similar ensemble spreads between the pre-industrial and historical simulations (Fig. 2d–f). On the other hand, hurricanes like Katrina, Irma and Maria are expected to significantly ($P = 0.05$) intensify with continued warming (Table 1 and Extended Data Fig. 3), corresponding to a shift towards greater intensities for the RCP8.5 simulations compared to the historical (Fig. 2d–f).

We extended the investigation to 15 tropical cyclone events sampled globally under three future climate scenarios, to address the robustness of the results. We performed the same analysis for all 15 tropical cyclones as was presented above for hurricanes Katrina, Irma and Maria, including an evaluation of the historical hindcast's ability to reproduce the observed tropical cyclone track. Of the 45 experiments, four were discarded for tropical cyclone tracks that deviated substantially from the historical case (Methods). Of the 15 tropical cyclones, 13 of which were intense, 11 show significant ($P = 0.05$) intensity increases, regardless of emissions scenario, with peak wind speed increases of 6–29 knots and minimum SLP reduced by 5–25 hPa (Table 1 and Extended Data Fig. 3). Changes are insignificant for hurricanes Andrew and Iniki, and hurricane Bob significantly weakens. Therefore, the experiments provide substantial support for strengthening of intense tropical cyclone events globally for the three future climate scenarios considered.

Finally, we quantified the uncertainty in the response of tropical cyclones to anthropogenic forcings owing to convective parameterization using simulations of hurricane Katrina at resolutions of 3 km, 9 km and 27 km. Regardless of resolution, these simulations produced insignificant changes in Katrina's intensity from the pre-industrial to historical climates, and a substantial and significant ($P = 0.05$) increase in intensity from the historical to RCP8.5 climates (Fig. 2d and Table 1), indicating that the qualitative simulated tropical cyclone response to

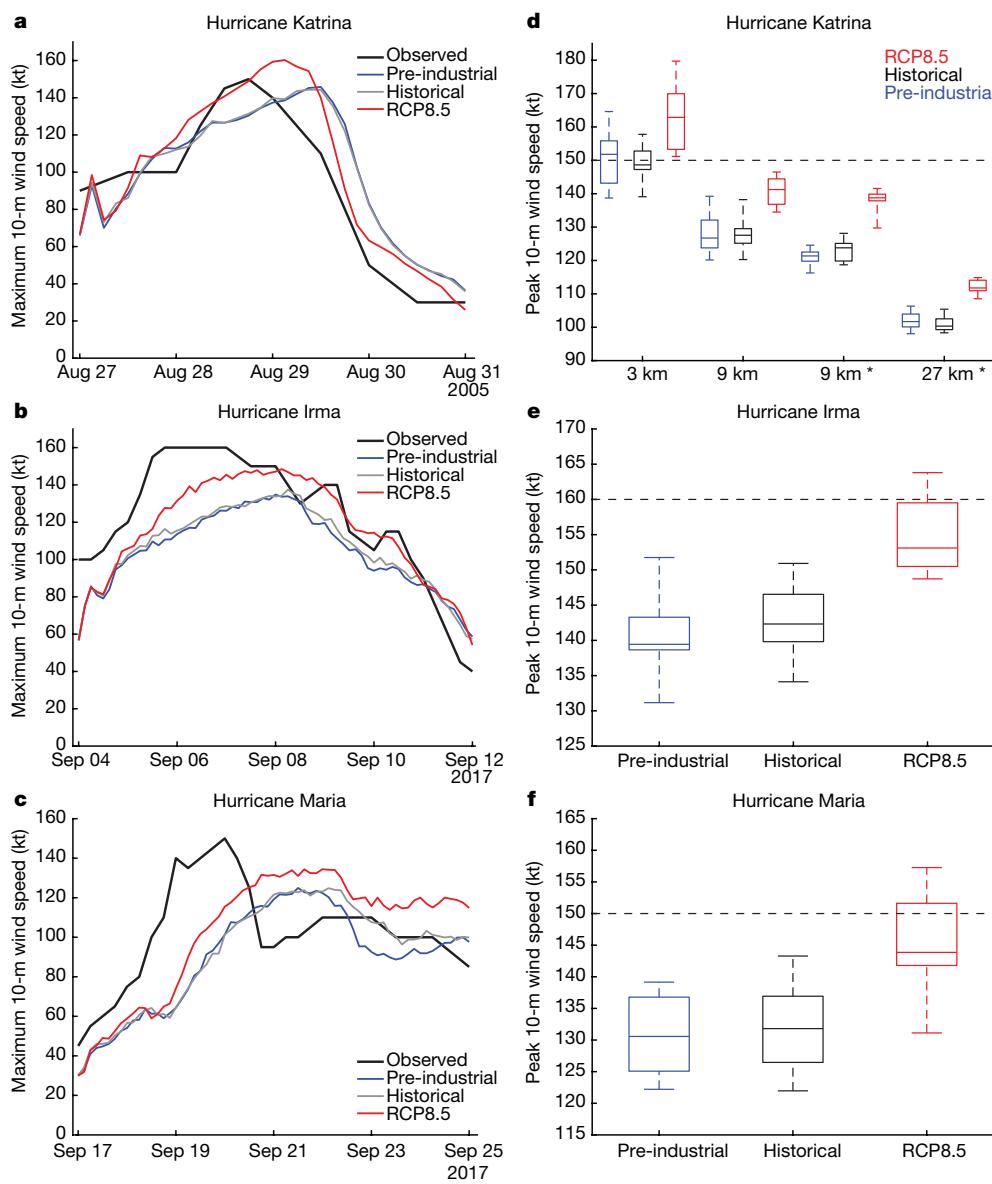


Fig. 2 | Time series and boxplots of tropical cyclone maximum 10-m wind speed. **a–c**, The time series of maximum 10-m wind speed (knots, kt) from observations (black) and the ensemble mean of the pre-industrial (blue), historical (grey) and RCP8.5 (red) simulations of hurricanes Katrina at 3-km resolution (**a**), Irma at 4.5-km resolution (**b**) and Maria at 4.5-km resolution (**c**). **d–f**, Boxplots of peak 10-m wind speed (kt) from the ten-member ensemble of pre-industrial (blue), historical (black) and

RCP8.5 (red) simulations of hurricane Katrina at 3-km, 9-km (with and without convective parameterization) and 27-km resolution (**d**), and of hurricanes Irma (**e**) and Maria (**f**) at 4.5-km resolution. The centre line denotes the median, box limits denote lower and upper quartiles, and whiskers denote the minimum and maximum. The observed peak intensity is marked with a horizontal dashed black line. Simulations that used convective parameterization are denoted by an asterisk.

anthropogenic forcing may be insensitive to use of convective parameterization and model resolution between 3 km and 27 km in this model, with additional work needed to make a generalized conclusion. Furthermore, the range of the future response is relatively small between resolutions, covering a 11–15 knot increase in maximum wind speed and a 11–14 hPa decrease in minimum SLP. However, model resolution substantially affects absolute intensity, as expected, with an ensemble-mean Category 5, Category 4 and Category 3 hurricane produced by the historical simulations at 3-km, 9-km and 27-km resolution, respectively.

Anthropogenic influences on tropical cyclone rainfall

Although tropical cyclone winds can cause substantial damages, heavy rainfall can pose an equal, if not greater, hazard. We analysed anthropogenic changes in rainfall within a reference frame centred on the tropical cyclone, called a ‘composite’ (Fig. 3), because even small changes in tropical cyclone track and translation speed confound a

geographically fixed analysis. The composites include the simulated tropical cyclone lifetime, excluding a generous 12-h spin-up, and cover ocean and land. Two levels of statistical significance ($P = 0.05$ and $P = 0.10$) are presented, as changes in rainfall tend to be noisy compared with wind speed and SLP. We found that climate change at the time of Katrina significantly ($P = 0.10$) enhanced rainfall rates by 4%–9% over an approximately $5^\circ \times 5^\circ$ box centred on the tropical cyclone, a result qualitatively insensitive to model resolution and use of convective parameterization (Table 2). Likewise, climate change at the time of hurricanes Irma and Maria significantly ($P = 0.10$) increased rainfall by 6% and 9%, respectively, but over a roughly $1.5^\circ \times 1.5^\circ$ box centred on the tropical cyclone (Table 2), owing to a concentration of the rainfall enhancements near the tropical cyclone centre (Fig. 3). Therefore, we find evidence that climate change so far has begun to enhance rainfall for these three tropical cyclones, with investigation of additional cases needed before making a general conclusion.

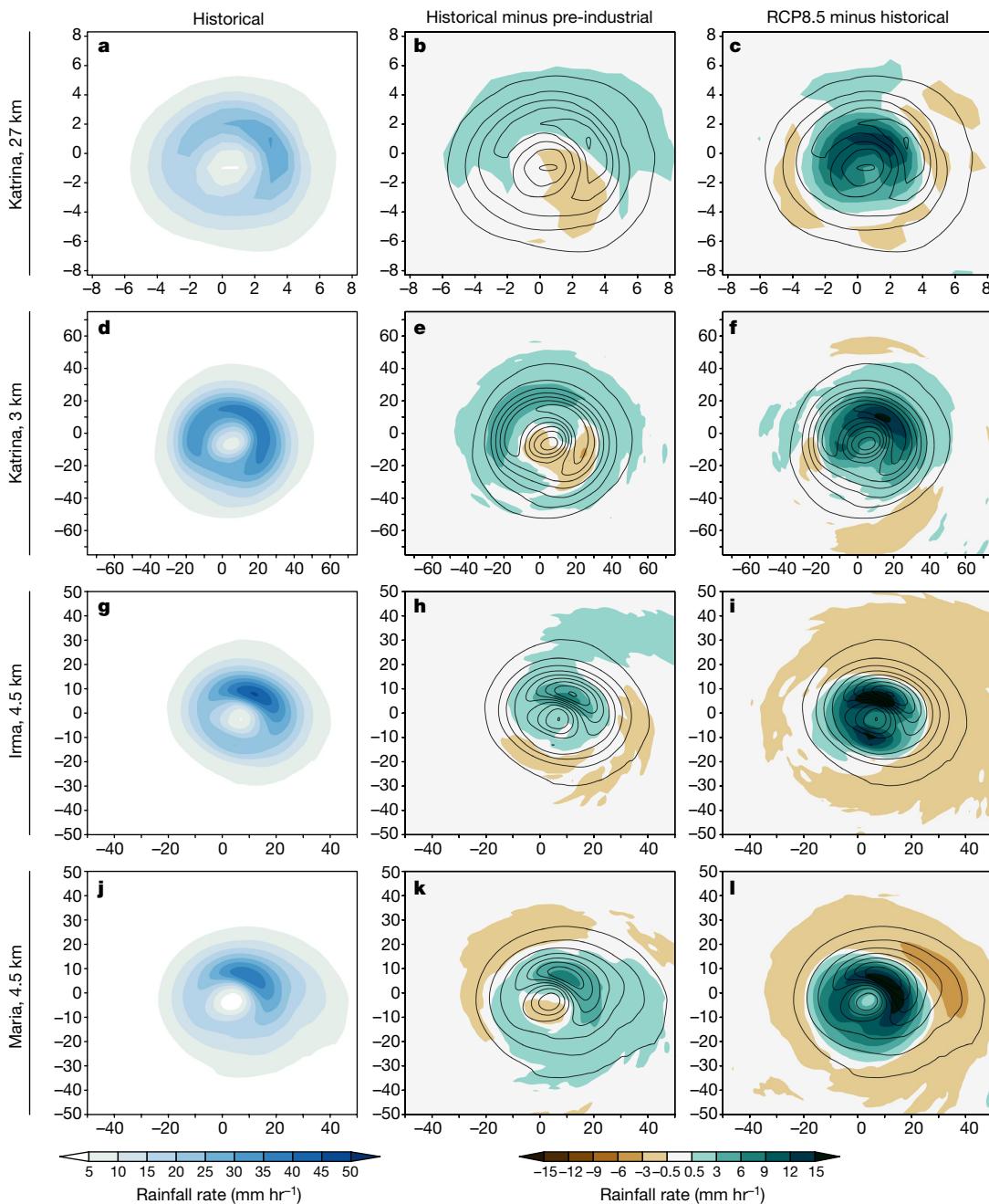


Fig. 3 | Tropical cyclone rainfall composites. a–c, Rainfall rate (colour scale) relative to tropical cyclone centre and throughout the simulated tropical cyclone lifetime from the ensemble mean of the historical (a), historical minus pre-industrial (b) and RCP8.5 minus historical (c) simulations of hurricane Katrina at 27-km resolution. d–l, As in a–c

In addition, we found robust increases in tropical cyclone rainfall with continued climate change along RCP4.5, RCP6.0 and RCP8.5 scenarios, which are significant for at least one Representative Concentration Pathway scenario for all 15 tropical cyclones except two ($P = 0.10$) or three ($P = 0.05$) (Table 2). The largest increases in rainfall tend to occur over the regions of heaviest historical rainfall (Fig. 3 and Extended Data Fig. 4). For some tropical cyclones, including hurricanes Irma and Maria (Fig. 3), there is a coherent spatial pattern in the future rainfall response characterized by drying in the outer-tropical-cyclone radii, resulting in rainfall responses that are stronger over an approximately $1.5^\circ \times 1.5^\circ$ area compared with a $5^\circ \times 5^\circ$ box (Table 2). Such outer-tropical-cyclone drying is not apparent or is weak for most tropical cyclones considered, including hurricanes Katrina, Floyd, Gafilo and Yasi (Fig. 3 and Extended Data Fig. 4). The future rainfall changes

but for simulations of hurricane Katrina at 3-km resolution (d–f) and of hurricanes Irma (g–i) and Maria (j–l) at 4.5-km resolution. Contours denote the rainfall rate (in millimetres per hour) from the corresponding historical simulation. The x and y axes show the number of model grid points from the tropical cyclone centre.

reach 25%–30% for some tropical cyclones under an RCP8.5 scenario (Table 2), exceeding what would be expected by Clausius–Clapeyron scaling alone given regional SST warming of about 2.5°C in these cases.

We next evaluated changes in extreme rainfall, which can be important for localized flooding, by considering the probability density functions of rainfall rates sampled three-hourly and including each model grid point within about $5^\circ \times 5^\circ$ centred on the tropical cyclone for the lifetime of the simulated storm (Fig. 4). The individual ensemble members of the 3-km-resolution hurricane Katrina simulations exhibit probabilities of extremely intense rainfall rates that consistently increase from the pre-industrial, to historical, to RCP8.5 experiments (Fig. 4a). This behaviour is also apparent in the ensemble means for simulations at 3-km and 27-km resolution; however, the coarser-resolution simulation consistently produces weaker extremes (Fig. 4b). The increasing

Table 2 | Changes in tropical cyclone rainfall

Basin	Tropical cyclone	Resolution	(Historical minus pre-industrial)/pre-industrial	(RCP4.5 minus historical)/historical	(RCP6.0 minus historical)/historical	(RCP8.5 minus historical)/historical
Atlantic	Katrina	27 km (P)	4.7**			13.0**
		9 km (P)	4.5*			12.7**
		9 km	5.0*			13.5**
		3 km	8.7**			14.4**
		4.5 km		7.1**	14.6**	16.5**
	Irma	4.5 km	4.2	4.5	8.8**	2.1
	Irma ^a	4.5 km	6.3*	17.5**	26.1**	27.8**
	Maria	4.5 km	4.4	7.0*	7.2*	7.7*
	Maria ^a	4.5 km	8.9**	21.8**	23.4**	36.9**
	Andrew	4.5 km		0.3	5.1	4.8
	Bob	4.5 km		6.5**	11.9**	13.5**
	Floyd	4.5 km		12.3**	13.5**	X
	Gilbert	4.5 km		13.5**	16.5**	25.3**
	Ike	4.5 km		15.0**	20.2**	26.5**
	Matthew	4.5 km		2.0	1.1	4.0
Eastern Pacific	Iniki	4.5 km	5.8*	4.9		15.2**
Northwest Pacific	Haiyan	4.5 km	9.5**	12.8**		31.3**
	Morakot	4.5 km	6.8*	X		X
	Songda	4.5 km		19.5**	10.6**	X
South Pacific	Yasi	4.5 km		15.6**	23.1**	35.2**
Southwest Indian	Gafilo	4.5 km		19.7**	16.8**	41.6**

The ensemble-mean change in rainfall is given (as a percentage) between the historical and pre-industrial simulations and between the RCP4.5, RCP6.0 and RCP8.5 simulations and the historical simulation, averaged over approximately $5^\circ \times 5^\circ$ and over $1.5^\circ \times 1.5^\circ$ (the latter is denoted ^a) boxes centred on the tropical cyclone. Cases of substantial differences between simulated and observed tropical cyclone tracks are denoted X (see Methods) and simulations that were not performed are blank. *Changes significant at the 10% level; **changes significant at the 5% level. Simulations that used convective parameterization are denoted 'P'.

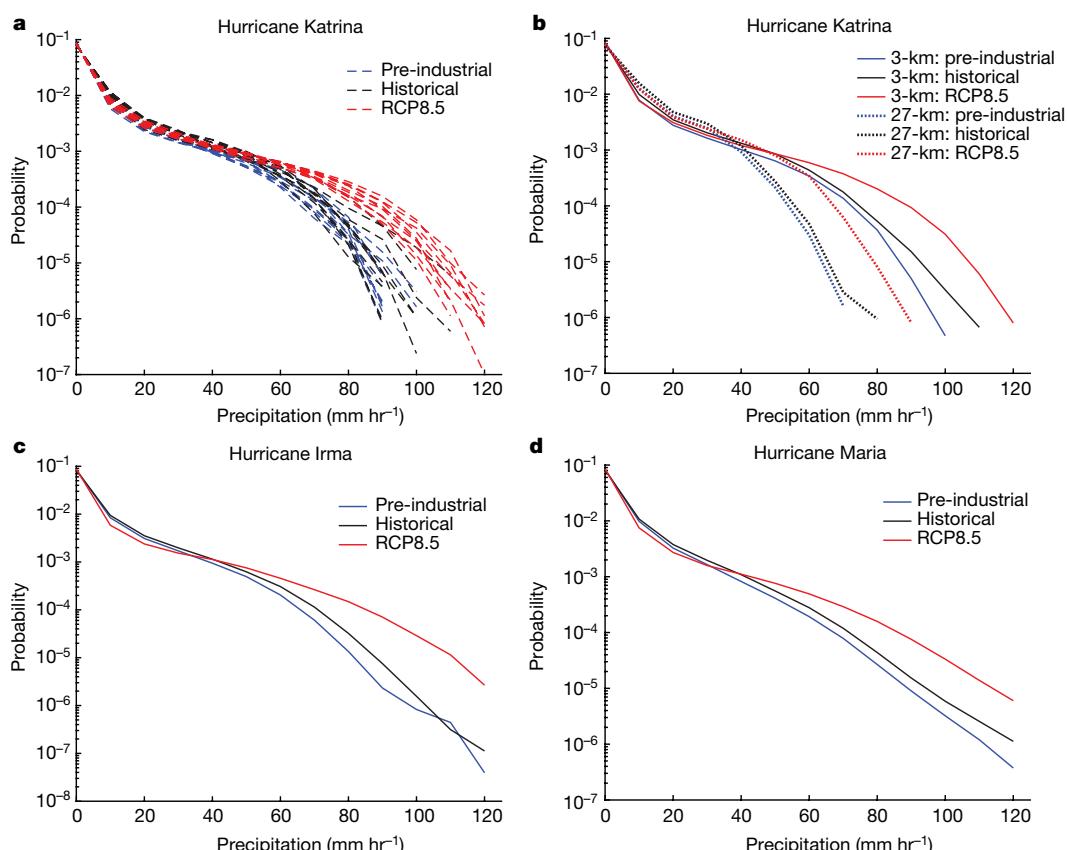


Fig. 4 | Probability density functions of tropical cyclone rainfall rates. **a–d**, Probability density functions of rainfall rates from each of ten ensemble members of the pre-industrial (blue), historical (black) and RCP8.5 (red) simulations of hurricane Katrina at 3-km resolution (a),

and from the ensemble means of simulations of hurricane Katrina at 3-km (solid lines) and 27-km (dotted lines) resolution (b) and of hurricanes Irma (c) and Maria (d) at 4.5-km resolution.

probability of extremely intense rainfall rates with anthropogenic warming is robust among hurricanes Katrina, Irma and Maria (Fig. 4).

Discussion

There is no consensus on whether climate change has yet affected tropical cyclone statistics, and how continued warming may influence many aspects of future tropical cyclone activity. We have improved our understanding of anthropogenic influences on tropical cyclones by quantifying how the intensity and rainfall of historically impactful tropical cyclone events could change if similar events occurred in cooler and warmer climates, using ten-member ensembles of convection-permitting hindcast simulations with boundary conditions adjusted to reflect the different climate states. We found that climate change so far has weakly and insignificantly influenced the wind speed and SLP-based intensities for hurricanes Katrina, Irma and Maria, suggesting the possibility that climate variability—rather than anthropogenic warming—may have driven the active 2005 and 2017 Atlantic hurricane seasons, which were indeed characterized by especially warm tropical Atlantic SSTs. However, climate change at the time of these hurricanes significantly enhanced rainfall by 4%–9% and increased the probability of extreme rainfall rates, suggesting that climate change to date has already begun to increase tropical cyclone rainfall. Investigation of additional tropical cyclones is needed before making a general conclusion.

We then considered how 15 tropical cyclone events sampled globally could change if similar events were to occur at the end of the twenty-first century under the RCP4.5, RCP6.0 and RCP8.5 scenarios. We found a substantial and significant future intensification in the majority (11 of 13) of intense tropical cyclone events, based on wind speed and SLP, consistent with maximum potential intensity theory. Analysis of SST and tropical tropopause temperature changes is planned to understand the physical mechanisms behind these responses. In addition, we found robust increases in future tropical cyclone rainfall, with some events exceeding what would be expected by Clausius–Clapeyron alone and some events demonstrating a spatial pattern with concentrated rainfall increases near the centre of the tropical cyclone and drying in the outer-tropical-cyclone radii. These future changes in tropical cyclone intensity and rainfall could exacerbate societal impacts associated with ocean wind-waves⁴⁵, storm surge, flooding, and forests and ecosystems⁴⁶. Simulations with and without convective parameterization suggest that convective parameterization introduces minimal uncertainty into the sign of projected changes in tropical cyclone intensity and rainfall, supporting confidence in projections of tropical cyclone activity from models with both parameterized convection and tropical-cyclone-permitting resolution (less than 0.25°).

The detection and attribution of anthropogenic changes in tropical cyclone events is a rapidly emerging science and methodology⁴⁷, especially as supercomputing advances enable ensembles of convection-permitting simulations. Our use of a dynamical climate model allows us to perform controlled experiments that focus on specific events and include various complexities of relevant physical processes. One important physical process for tropical cyclones that is missing from our model design is atmosphere–ocean coupling. In reality, tropical cyclone winds typically induce a ‘cold wake’ of upper-ocean temperatures that can provide a negative feedback on tropical cyclone intensity, depending on the tropical cyclone’s intensity and translation speed and the ocean heat content and salinity structure^{40,48,49}. Therefore, lack of coupling in the model can lead to tropical cyclones that are more intense and frequent compared to slab–ocean and fully coupled atmosphere–ocean simulations^{50,51}. The atmosphere-only simulations presented in this study may overestimate tropical cyclone intensity, and additional research would help to quantify this uncertainty. In addition, because we used a single climate model, we have not examined model structural uncertainty, and results from other convection-permitting models could vary from those presented here.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0673-2>.

Received: 15 April 2018; Accepted: 4 September 2018;
Published online 14 November 2018.

1. National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Information (NCEI). *Billion-Dollar Weather and Climate Disasters* <https://www.ncdc.noaa.gov/billions/> (NOAA NCEI, 2018).
2. Grossmann, I. & Morgan, M. G. Tropical cyclones, climate change, and scientific uncertainty: what do we know, what does it mean, and what should be done? *Clim. Change* **108**, 543–579 (2011).
3. Walsh, K. J. E. et al. Tropical cyclones and climate change. *WIREs Clim. Chang.* **7**, 65–89 (2016).
4. Sobel, A. H. et al. Human influence on tropical cyclone intensity. *Science* **353**, 242–246 (2016).
5. Landsea, C. W., Vecchi, G. A., Bengtsson, L. & Knutson, T. R. Impact of duration thresholds on Atlantic tropical cyclone counts. *J. Clim.* **23**, 2508–2519 (2010).
6. Landsea, C. W., Harper, B. A., Hoarau, K. & Knaff, J. A. Can we detect trends in extreme tropical cyclones? *Science* **313**, 452–454 (2006).
7. Vecchi, G. A. & Knutson, T. R. On estimates of historical north Atlantic tropical cyclone activity. *J. Clim.* **21**, 3580–3600 (2008).
8. Emanuel, K. Increasing destructiveness of tropical cyclones over the past 30 years. *Nature* **436**, 686–688 (2005).
9. Webster, P. J., Holland, G. J., Curry, J. A. & Chang, H. R. Changes in tropical cyclone number, duration, and intensity in a warming environment. *Science* **309**, 1844–1846 (2005).
10. Klotzbach, P. J. Trends in global tropical cyclone activity over the past twenty years (1986–2005). *Geophys. Res. Lett.* **33**, L10805 (2006).
11. Kossin, J. P., Knapp, K. R., Vimont, D. J., Murnane, R. J. & Harper, B. A. A globally consistent reanalysis of hurricane variability and trends. *Geophys. Res. Lett.* **34**, L04815 (2007).
12. Landsea, C. W., Pielke, R. A., Mestas-Nunez, A. & Knaff, J. A. Atlantic basin hurricanes: indices of climatic changes. *Clim. Change* **42**, 89–129 (1999).
13. Goldenberg, S. B., Landsea, C. W., Mestas-Nunez, A. M. & Gray, W. M. The recent increase in Atlantic hurricane activity: causes and implications. *Science* **293**, 474–479 (2001).
14. Gray, W. M. Atlantic seasonal hurricane frequency. 1. El-Nino and 30-mb quasi-biennial oscillation influences. *Mon. Weath. Rev.* **112**, 1649–1668 (1984).
15. Vimont, D. J. & Kossin, J. P. The Atlantic Meridional Mode and hurricane activity. *Geophys. Res. Lett.* **34**, L07709 (2007).
16. Patricola, C. M., Saravanan, R. & Chang, P. The impact of the El Nino–Southern Oscillation and Atlantic Meridional Mode on seasonal Atlantic tropical cyclone activity. *J. Clim.* **27**, 5311–5328 (2014).
17. Patricola, C. M., Chang, P. & Saravanan, R. Degree of simulated suppression of Atlantic tropical cyclones modulated by flavour of El Nino. *Nat. Geosci.* **9**, 155–160 (2016).
18. Gualdi, S., Scoccimarro, E. & Navarra, A. Changes in tropical cyclone activity due to global warming: results from a high-resolution coupled general circulation model. *J. Clim.* **21**, 5204–5228 (2008).
19. Knutson, T. R., Sirutis, J. J., Garner, S. T., Vecchi, G. A. & Held, I. M. Simulated reduction in Atlantic hurricane frequency under twenty-first-century warming conditions. *Nat. Geosci.* **1**, 359–364 (2008).
20. Knutson, T. R. et al. Tropical cyclones and climate change. *Nat. Geosci.* **3**, 157–163 (2010).
21. Wehner, M. et al. Resolution dependence of future tropical cyclone projections of CAM5.1 in the US CLIVAR Hurricane Working Group idealized configurations. *J. Clim.* **28**, 3905–3925 (2015).
22. Wehner, M. F., Reed, K. A., Loring, B., Stone, D. & Krishnan, H. Changes in tropical cyclones under stabilized 1.5°C and 2.0°C global warming scenarios as simulated by the Community Atmospheric Model under the HAPPI protocols. *Earth Syst. Dyn.* **9**, 187–195 (2018).
23. Emanuel, K. A. Downscaling CMIP5 climate models shows increased tropical cyclone activity over the 21st century. *Proc. Natl. Acad. Sci. USA* **110**, 12219–12224 (2013).
24. Emanuel, K. A. The dependence of hurricane intensity on climate. *Nature* **326**, 483–485 (1987).
25. Knutson, T. R. & Tuleya, R. E. Impact of CO₂-induced warming on simulated hurricane intensity and precipitation: sensitivity to the choice of climate model and convective parameterization. *J. Clim.* **17**, 3477–3495 (2004).
26. Bender, M. A. et al. Modeled impact of anthropogenic warming on the frequency of intense Atlantic hurricanes. *Science* **327**, 454–458 (2010).
27. Hill, K. A. & Lackmann, G. M. The impact of future climate change on TC intensity and structure: a downscaling approach. *J. Clim.* **24**, 4644–4661 (2011).
28. Knutson, T. R. et al. Dynamical downscaling projections of twenty-first-century Atlantic hurricane activity: CMIP3 and CMIP5 model-based scenarios. *J. Clim.* **26**, 6591–6617 (2013).
29. Walsh, K. J. E. et al. Hurricanes and climate: the U.S. CLIVAR Working Group on hurricanes. *Bull. Am. Meteorol. Soc.* **96**, 997–1017 (2015).
30. Villarini, G. et al. Sensitivity of tropical cyclone rainfall to idealized global-scale forcings. *J. Clim.* **27**, 4622–4641 (2014).
31. Scoccimarro, E. et al. Intense precipitation events associated with landfalling tropical cyclones in response to a warmer climate and increased CO₂. *J. Clim.* **27**, 4642–4654 (2014).

32. Wright, D. B., Knutson, T. R. & Smith, J. A. Regional climate model projections of rainfall from US landfalling tropical cyclones. *Clim. Dyn.* **45**, 3365–3379 (2015).

33. Scoccimarro, E. et al. in *Hurricanes and Climate Change* (eds Collins, J. & Walsh, K.) 243–255 (Springer, Cham, 2017).

34. Allen, M. R. & Ingram, W. J. Constraints on future changes in climate and the hydrologic cycle. *Nature* **419**, 224–232 (2002).

35. Risser, M. D. & Wehner, M. F. Attributable human-induced changes in the likelihood and magnitude of the observed extreme precipitation during hurricane Harvey. *Geophys. Res. Lett.* **44**, 12457–12464 (2017).

36. van Oldenborgh, G. J. et al. Attribution of extreme rainfall from hurricane Harvey, August 2017. *Environ. Res. Lett.* **12**, 124009 (2017).

37. Wang, S. Y. et al. Quantitative attribution of climate effects on hurricane Harvey's extreme rainfall in Texas. *Environ. Res. Lett.* **13**, 054014 (2018).

38. Emanuel, K. Assessing the present and future probability of hurricane Harvey's rainfall. *Proc. Natl Acad. Sci. USA* **114**, 12681–12684 (2017).

39. Gray, W. M. Global view of origin of tropical disturbances and storms. *Mon. Weath. Rev.* **96**, 669–700 (1968).

40. Huang, P., Lin, I. I., Chou, C. & Huang, R. H. Change in ocean subsurface environment to suppress tropical cyclone intensification under global warming. *Nat. Commun.* **6**, 7188 (2015).

41. Emanuel, K., Solomon, S., Folini, D., Davis, S. & Cagnazzo, C. Influence of tropical tropopause layer cooling on Atlantic hurricane activity. *J. Clim.* **26**, 2288–2301 (2013).

42. Wing, A. A., Emanuel, K. & Solomon, S. On the factors affecting trends and variability in tropical cyclone potential intensity. *Geophys. Res. Lett.* **42**, 8669–8677 (2015).

43. Vecchi, G. A. & Soden, B. J. Increased tropical Atlantic wind shear in model projections of global warming. *Geophys. Res. Lett.* **34**, L08702 (2007).

44. Kaplan, J. et al. *Improvement in the Rapid Intensity Index by Incorporation of Inner-core Information*. JHT final report. https://www.nhc.noaa.gov/jht/09-11reports/final_Kaplan_JHT11.pdf (NOAA, 2011).

45. Timmermans, B., Patricola, C. M. & Wehner, M. F. Simulation and analysis of hurricane-driven extreme wave climate under two ocean warming scenarios. *Oceanography* **31**, <https://doi.org/10.5670/oceanog.2018.218> (2018).

46. Feng, Y. et al. Rapid remote sensing assessment of impacts from hurricane Maria on forests of Puerto Rico. Preprint at <https://peerj.com/preprints/26597/> (2018).

47. Wehner, M. F., Zarzycki, C. & Patricola, C. M. in *Hurricane Risk* (eds Collins, J. & Walsh, K.) Ch. 12 (Springer, Cham, in the press).

48. Lin, I. I., Pun, I. F. & Wu, C. C. Upper-ocean thermal structure and the western north Pacific category 5 typhoons. Part II: dependence on translation speed. *Mon. Weath. Rev.* **137**, 3744–3757 (2009).

49. Balaguru, K. et al. Ocean barrier layers' effect on tropical cyclone intensification. *Proc. Natl Acad. Sci. USA* **109**, 14343–14347 (2012).

50. Zarzycki, C. M. Tropical cyclone intensity errors associated with lack of two-way ocean coupling in high-resolution global simulations. *J. Clim.* **29**, 8589–8610 (2016).

51. Li, H. & Srivastava, R. L. Tropical cyclone activity in the high-resolution community earth system model and the impact of ocean coupling. *J. Adv. Model. Earth Syst.* **10**, 165–186 (2018).

Acknowledgements This material is based on work supported by the US Department of Energy, Office of Science, Office of Biological and Environmental Research, Climate and Environmental Sciences Division, Regional and Global Climate Modeling Program, under award number DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the US Department of Energy under contract number DE-AC02-05CH11231. We thank H. Krishnan for setting up access to the simulation data at NERSC.

Reviewer information *Nature* thanks J. Manganello and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.M.P. and M.F.W. conceived the project and developed the methodology. C.M.P. performed the simulations, with climate perturbations from M.F.W., and analysed the data. C.M.P. wrote the manuscript with contributions from M.F.W.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0673-2>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to C.M.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

We performed hindcast simulations with the Weather Research and Forecasting (WRF) regional climate model⁵² version 3.8.1, which is developed by the National Center for Atmospheric Research (NCAR). The regional model is well suited for this study for several reasons. First, the use of lateral boundary conditions (LBCs) allows us to prescribe a tighter constraint on the large-scale circulation (that is, steering flow) of the tropical cyclone hindcast than if a global model were used. This is beneficial because it is necessary for the hindcasts to reproduce observed tropical cyclone tracks well, given that tropical cyclone characteristics such as intensity and rainfall are sensitive to underlying SST and surrounding environmental conditions. In addition, such ‘well behaved’ tracks among different climate scenarios enables a ‘fair’ comparison of the tropical cyclone responses. That is, a simulated tropical cyclone that deviates substantially from the observed track does not truly represent that tropical cyclone. (We typically used a criterion of about 3° of latitude or longitude, with some subjective judgement.) Second, whereas global climate models typically use the hydrostatic approximation to simplify the vertical momentum equation, WRF is non-hydrostatic and therefore more appropriate for simulating small-scale convective processes. Finally, the regional domain allows us to perform ensembles of simulations at convection-permitting resolution, which would be computationally less feasible with a global model.

The control simulations consist of hindcasts representing 15 tropical cyclone events (Fig. 1 and Extended Data Table 1) in the historical conditions in which they actually occurred. We selected tropical cyclones that were particularly impactful and represent various tropical cyclone basins. The North Indian Ocean was omitted owing to model instability probably associated with the Tibetan Plateau, and hurricane Harvey was omitted owing to poor hindcast skill. Initial conditions and LBCs for the historical hindcast simulations were taken from the six-hourly National Centers for Environmental Prediction (NCEP) Climate Forecast System (CFS) Reanalysis⁵³ for all tropical cyclones occurring before March 2011, and from the NCEP Climate Forecast System Version 2 (ref. ⁵⁴) for tropical cyclones occurring in March 2011 or later. No adjustments or data assimilation were performed on the initial conditions or LBCs. Model initialization time (Extended Data Table 1) was chosen to represent the tropical cyclone for as much of its lifetime as possible, while still being able to realistically simulate the observed track, since an earlier initialization time generally produced larger deviations between the observed and simulated tropical cyclone track. The tropical cyclone intensity within the model adjusts from its initial condition within hours. We did not test whether the simulated anthropogenic influence on tropical cyclones is sensitive to initialization time. SST was prescribed from the daily 0.25° National Oceanic and Atmospheric Administration Optimum Interpolation (NOAA-OI) dataset⁵⁵ for all tropical cyclones, except hurricanes Irma and Maria, which used the NCEP Climate Forecast System Version 2. Greenhouse gas concentrations, including CO₂, CH₄, N₂O, CFC-11, CFC-12 and CCl₄, were prescribed according to refs ^{56,57}. A ten-member ensemble of each simulation was generated using the Stochastic Kinetic Energy Backscatter Scheme (SKEBS)⁵⁸, which represents uncertainty from interactions with unresolved scales by introducing temporally and spatially correlated perturbations to the rotational wind components and potential temperature. The SST, initial condition and LBCs are identical for each ensemble member within a simulation set.

We also performed experiments representing hurricanes Katrina, Irma and Maria if they were to occur in a pre-industrial climate and all 15 tropical cyclone events at the end of the twenty-first century under the RCP4.5, RCP6.0 and RCP8.5 emissions scenarios, as permitted by supercomputing resources. SSTs, initial conditions and LBCs for the pre-industrial and Representative Concentration Pathway experiments were based on those from the historical simulations, with adjustments to remove and add, respectively, the thermodynamic component of anthropogenic climate change, using the 'pseudo-global warming' approach detailed in refs^{47,59}. In the pseudo-global warming experiments, the model's boundary conditions use the same input data as in the control simulations for the historical period, but with a climate change signal added. This methodology has been used to study anthropogenic influences on individual tropical cyclone events at similar horizontal resolutions used in this study^{60–65}. The novelty here is in investigating over a dozen tropical cyclone cases under multiple emissions scenarios at such a resolution. The variables adjusted in the LBCs include temperature, relative humidity and geopotential height. We did not adjust horizontal winds in the LBCs to minimize possible perturbations to the simulated hurricane track, although tests on a subset of simulations showed that the response in tropical cyclone intensity to anthropogenic forcing is insensitive to whether circulation changes were applied to the LBCs. The experimental design, therefore, prescribes no changes in large-scale vertical wind shear. We note that any potential changes in vertical wind shear⁴³ may be expected to change the summary statistics of tropical cyclone activity (for example, average annual number of tropical cyclones). However, even given average changes in wind shear, it is conceivable that individual tropical cyclone events may occur under shear conditions similar to those of the present climate, especially since some climate models project relatively weak shear changes in the Atlantic

and Pacific basins⁶⁶. Therefore, by prescribing zero change in horizontal winds in the climate change simulations, the large-scale vertical shear state is included in the conditionalities of the 'worst-case-scenario' tropical cyclone event occurrence. This allows us to evaluate changes in tropical cyclone magnitudes given similar shear conditions, which may become more or less likely in changing climates.

The variables adjusted in the initial conditions include surface temperature (land and sea), 2-m air temperature, 2-m specific humidity, SLP and surface pressure. Greenhouse gas concentrations were modified in the WRF climate model according to refs^{56,57,67}. The experimental design is similar to the hindcast methodology used to understand anthropogenic contributions to the extreme flood event that affected the Boulder, Colorado, region in September 2013⁶⁸.

Anthropogenic climate change from the pre-industrial to historical period was estimated using Community Atmosphere Model (CAM) simulations from the Climate of the 20th Century Plus Detection and Attribution (C20C+ D&A) Project⁶⁹ (D. A. Stone et al., submitted manuscript). The ‘factual’ C20C+ simulation consists of a 50-member ensemble of 1° resolution CAM5.1 integrations forced with historical radiative and land-surface boundary conditions and SST, and the ‘counterfactual’ simulation uses radiative forcing from the year 1855, with SST and sea ice modified using perturbations from coupled atmosphere–ocean simulations of the Coupled Model Intercomparison Project Phase 5 (CMIP5; D. A. Stone & P. Pall, submitted manuscript). The climate change perturbation for the pre-industrial hurricane Katrina experiment was calculated as the difference between the factual and the counterfactual C20C+ simulations for August 2005; this perturbation was then subtracted from the historical boundary conditions. For hurricanes Irma and Maria, the perturbation was estimated as the difference between the September 1996–2016 climatology of the factual minus counterfactual C20C+ simulations, as the C20C+ simulations did not extend to 2017 at the time of this study.

Anthropogenic climate change for the end of the twenty-first century was based on simulations from the Community Climate System Model (CCSM4) of the CMIP5. The climate change perturbation for the RCP8.5 hurricane Katrina experiment was calculated as the 2081–2100 August climatology from the CCSM4 RCP8.5 simulation minus the 1980–2000 August climatology from the CCSM4 historical simulation. This perturbation was then added to the historical boundary conditions. The perturbations for all other tropical cyclones were calculated in the same way, but for the month in which the tropical cyclone occurred (for example, September for hurricanes Irma and Maria).

By using one global model to provide climate change perturbations, the results here apply for the climate sensitivity characteristic of that model. The uncertainty due to the range of climate sensitivities among different models was not accounted for, in favour of using supercomputing resources towards 15 tropical cyclone events, convection-permitting resolution, ten-member ensembles and multiple Representative Concentration Pathway scenarios. We note that the climate sensitivity of the CCSM4 model is among the lower of the coupled atmosphere-ocean global climate models of CMIP5^{70,71}, suggesting that the estimates of future change provided by this study may be conservative. The SST forcings for the CAM simulations from the C20C+ D&A Project were based on the multi-model mean of the CMIP5, suggesting that the estimates of climate change influences from pre-industrial to present are near the centre of the range of models.

Simulations of all tropical cyclone events were performed at a convection-permitting horizontal resolution of 4.5 km, with 44 levels in the vertical and a model top at 20 hPa. To investigate uncertainty in the response of tropical cyclones to anthropogenic forcings due to convective parameterization, we performed additional simulations of Hurricane Katrina at horizontal resolutions of 3 km without parameterization, 9 km both without and with parameterization (Kain-Fritsch), and 27 km with parameterization, with 35 levels in the vertical and a model top at 50 hPa. The results are insensitive to vertical resolution and model top choices.

Simulated tropical cyclone coordinates are defined using the location of minimum SLP. Simulated three-hourly instantaneous maximum 10-m tropical cyclone wind speeds are compared with the observed six-hourly maximum 1-min average sustained 10-m wind speed from the Revised Hurricane Database (HURDAT2^{72,73}) and the Joint Typhoon Warning Center (JTWC) dataset as archived in the International Best Track Archive for Climate Stewardship (IBTrACS⁷⁴) v03r10 database. Such differences in maximum wind speed definitions generate uncertainty in comparisons between observations and model simulations, and it is unclear whether there is a tendency for one definition to be systematically biased in a particular direction. The historical simulations appear to produce tropical cyclones with slightly weaker intensities than observed (Fig. 1), which may be related to these differences in intensity definitions between the model and observations, or to model limitations in horizontal resolution and/or physical approximations. Despite this uncertainty, the simulations with convection-permitting resolution perform substantially better in reproducing the approximate tropical cyclone intensities than simulations with convective parameterization (Fig. 3d). In addition, we acknowledge that while climate models are imperfect, the robust climate change response for hurricane Katrina at horizontal resolutions between

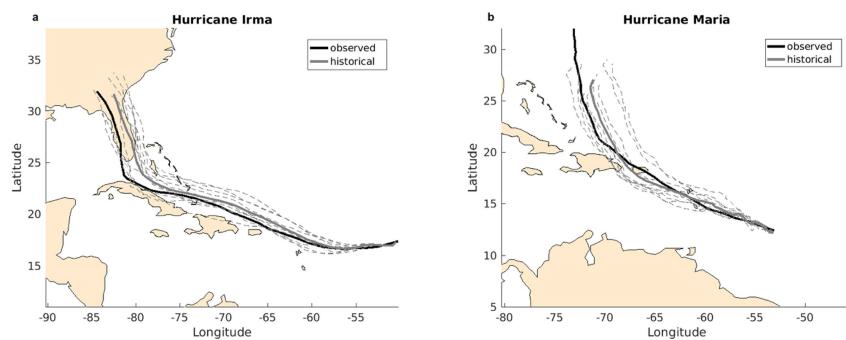
3 km and 27 km provides support for the contention that 4.5-km resolution is sufficient to capture the influence of climate change on tropical cyclones in the full set of experiments.

Code availability. Code for the WRF model, version 3.8.1, is available at <http://www2.mmm.ucar.edu/wrf/users/downloads.html>. Analytical scripts are available from the corresponding author on request.

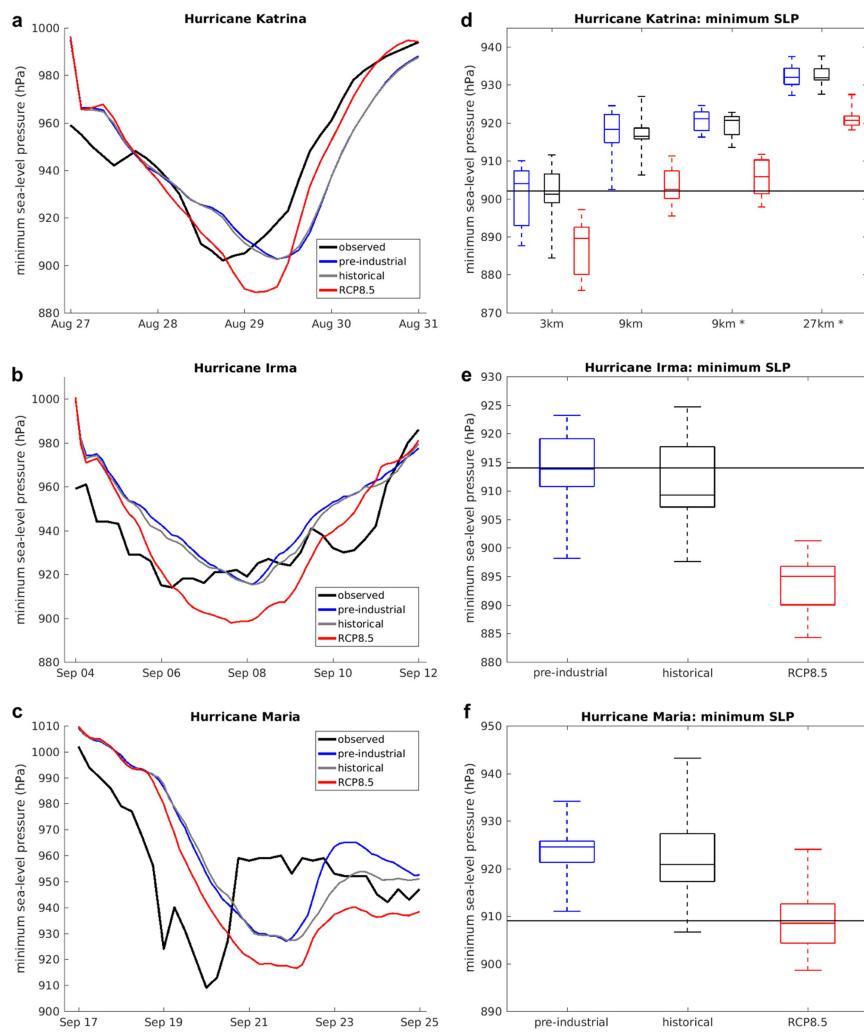
Data availability

Simulation data are available at the National Energy Research Scientific Computing Center (NERSC) at <http://portal.nersc.gov/cascade/TC/>.

52. Skamarock, W. C. & Klemp, J. B. A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *J. Comput. Phys.* **227**, 3465–3485 (2008).
53. Saha, S. et al. The NCEP Climate Forecast System Reanalysis. *Bull. Am. Meteorol. Soc.* **91**, 1015–1058 (2010).
54. Saha, S. et al. The NCEP Climate Forecast System Version 2. *J. Clim.* **27**, 2185–2208 (2014).
55. Reynolds, R. W. et al. Daily high-resolution-blended analyses for sea surface temperature. *J. Clim.* **20**, 5473–5496 (2007).
56. Tsutsumi, Y., Mori, K., Hihara, T., Ikegami, M. & Conway, T. J. *Technical Report of Global Analysis Method for Major Greenhouse Gases by the World Data Center for Greenhouse Gases*. GAW Report No. 184, https://www.wmo.int/pages/prog/arep/gaw/documents/TD_1473_GAW184_web.pdf (World Meteorological Organization, 2009).
57. Bullister, J. L. *Atmospheric Histories (1765–2015) for CFC-11, CFC-12, CFC-113, CCl4, SF6 and N2O*. NDP-095. http://cdiac.ess-dive.lbl.gov/ftp/oceans/CFC_ATM_Hist/CFC_ATM_Hist_2015/ (Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy, 2015).
58. Shutts, G. A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Q. J. R. Meteorol. Soc.* **131**, 3079–3102 (2005).
59. Schär, C., Frei, C., Lüthi, D. & Davies, H. C. Surrogate climate-change scenarios for regional climate models. *Geophys. Res. Lett.* **23**, 669–672 (1996).
60. Takayabu, I. et al. Climate change effects on the worst-case storm surge: a case study of typhoon Haiyan. *Environ. Res. Lett.* **10**, 064011 (2015).
61. Lackmann, G. M. Hurricane Sandy before 1900 and after 2100. *Bull. Am. Meteorol. Soc.* **96**, 547–560 (2015).
62. Ito, R., Takemi, T. & Arakawa, O. A possible reduction in the severity of typhoon wind in the northern part of Japan under global warming: a case study. *Sci. Online Lett. Atmos.* **12**, 100–105 (2016).
63. Nakamura, R., Shibayama, T., Esteban, M. & Iwamoto, T. Future typhoon and storm surges under different global warming scenarios: case study of typhoon Haiyan (2013). *Natural Hazards* **82**, 1645–1681 (2016).
64. Takemi, T., Ito, R. & Arakawa, O. Effects of global warming on the impacts of Typhoon Mireille (1991) in the Kyushu and Tohoku regions. *Hydrolog. Res. Lett.* **10**, 81–87 (2016).
65. Kanada, S. et al. A multimodel intercomparison of an intense typhoon in future, warmer climates by four 5-km-mesh models. *J. Clim.* **30**, 6017–6036 (2017).
66. Wehner, M. F. et al. Towards direct simulation of future tropical cyclone statistics in a high-resolution global atmospheric model. *Adv. Meteorol.* **2010**, 915303 (2010).
67. Meinshausen, M. et al. The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Clim. Change* **109**, 213–241 (2011).
68. Pall, P. et al. Diagnosing conditional anthropogenic contributions to heavy Colorado rainfall in September 2013. *Weather Clim. Extrem.* **17**, 1–6 (2017).
69. Stone, D. A. et al. A basis set for exploration of sensitivity to prescribed ocean conditions for estimating human contributions to extreme weather in CAM5.1-1degree. *Weather Clim. Extrem.* **19**, 10–19 (2018).
70. Vial, J., Dufresne, J.-L. & Bony, S. On the interpretation of inter-model spread in CMIP5 climate sensitivity estimates. *Clim. Dyn.* **41**, 3339–3362 (2013).
71. Andrews, T., Gregory, J. M., Webb, M. J. & Taylor, K. E. Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models. *Geophys. Res. Lett.* **39**, L09712 (2012).
72. Landsea, C. W. et al. in *Hurricanes and Typhoons: Past, Present and Future* (eds Murnane, R. J. & Liu, K.-B.) 177–221 (Columbia Univ. Press, New York, 2004).
73. Landsea, C. W. & Franklin, J. L. Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Weath. Rev.* **141**, 3576–3592 (2013).
74. Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J. & Neumann, C. J. The International Best Track Archive for Climate Stewardship (IBTrACS): unifying tropical cyclone best track data. *Bull. Am. Meteorol. Soc.* **91**, 363–376 (2010).

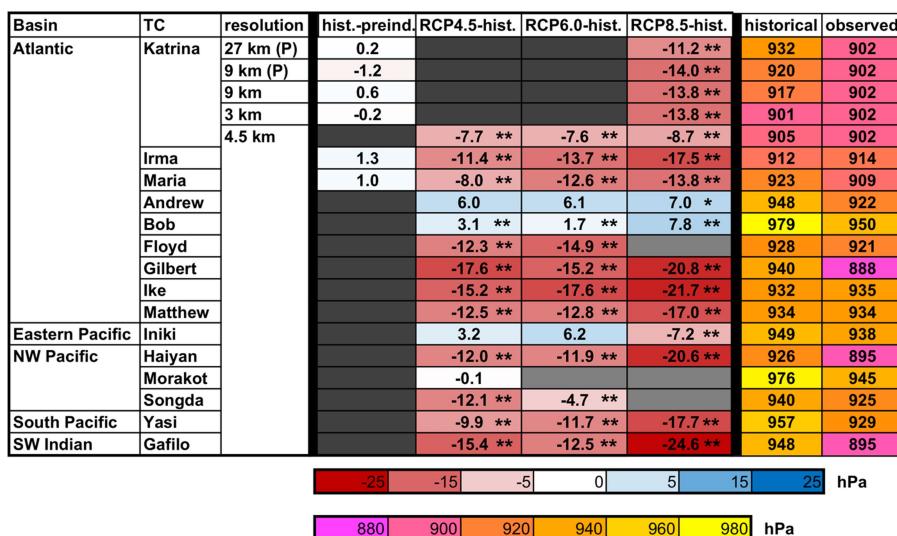


Extended Data Fig. 1 | Tropical cyclone tracks. **a, b,** The observed hurricane track (black) with simulated tropical cyclone tracks from ten ensemble members (grey dashed lines) and the ensemble mean (grey solid line) of the historical simulation for hurricanes Irma (a) and Maria (b) at 4.5-km resolution.



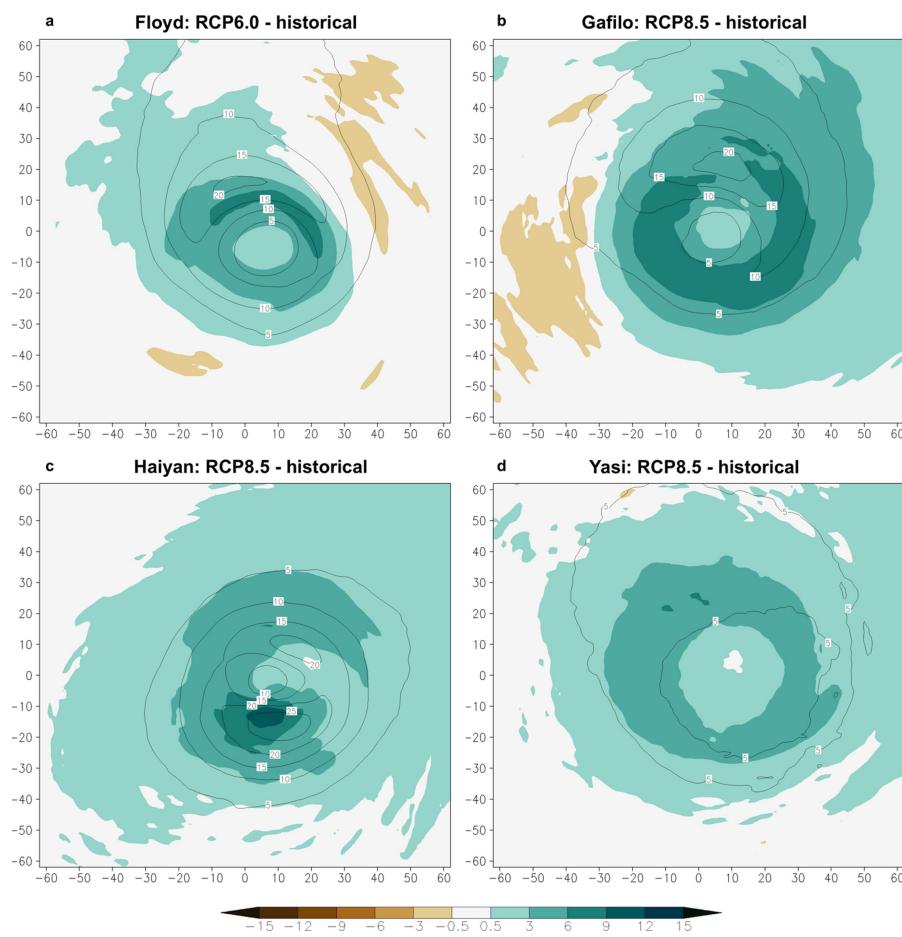
Extended Data Fig. 2 | Time series and boxplots of tropical cyclone minimum SLP. **a–c.** The time series of minimum SLP (hPa) from observations (black) and the ensemble mean of the pre-industrial (blue), historical (grey) and RCP8.5 (red) simulations of hurricane Katrina at 3-km resolution (**a**) and hurricanes Irma (**b**) and Maria (**c**) at 4.5-km resolution. **d–f.** Boxplots of minimum SLP (hPa) from the ten-member ensemble of pre-industrial (blue), historical (black) and RCP8.5 (red)

simulations of hurricane Katrina at 3-km, 9-km and 27-km resolution (**d**), and of hurricanes Irma (**e**) and Maria (**f**) at 4.5-km resolution. The centre line denotes the median, box limits denote lower and upper quartiles, and whiskers denote the minimum and maximum. The observed minimum SLP is marked with a horizontal black line. Simulations that used convective parameterization are denoted by asterisks.



Extended Data Fig. 3 | Tropical cyclone minimum SLP. Heatmaps are shown of the ensemble-mean difference in minimum SLP (in hPa) between the historical and pre-industrial simulations and between the RCP4.5, RCP6.0 and RCP8.5 simulations and the historical simulation (blue/red), with minimum SLP from observations and the ensemble-mean historical

simulation (yellow/magenta). Light grey denotes substantial differences between the simulated and the observed tropical cyclone tracks and dark grey denotes simulations that were not performed. *Changes significant at the 10% level; **changes significant at the 5% level. Simulations that used convective parameterization are denoted 'P'.


Extended Data Fig. 4 | Tropical cyclone rainfall composites.

a–d, Rainfall rate (colour scale, in millimetres per hour) relative to the tropical cyclone centre and throughout the simulated tropical cyclone lifetime from the ensemble mean of the RCP6.0 minus historical simulation of hurricane Floyd (**a**) and the RCP8.5 minus historical

simulation of cyclone Gafilo (**b**), typhoon Haiyan (**c**) and cyclone Yasi (**d**) at 4.5-km resolution. Contours denote the rainfall rate (in millimetres per hour) from the corresponding historical simulation. The axes show the number of model grid points from the tropical cyclone centre.

Extended Data Table 1 | List of tropical cyclone events

Basin	TC	Historical simulation period
Atlantic	Katrina	27 – 31 Aug, 2005
	Irma	4 – 13 Sep, 2017
	Maria	17 – 25 Sep, 2017
	Andrew	23 – 28 Aug, 1992
	Bob	18 – 21 Aug, 1991
	Floyd	13 – 18 Sep, 1999
	Gilbert	13 – 19 Sep, 1988
	Ike	6 – 15 Sep, 2008
	Matthew	1 – 7 Oct, 2016
East Pacific	Iniki	8 – 14 Sep, 1992
North West Pacific	Haiyan	5 – 11 Nov, 2013
	Morakot	6 – 12 Aug, 2009
	Songda	3 – 9 Sep, 2004
South Pacific	Yasi	31 Jan – 5 Feb, 2011
South West Indian	Gafilo	4 – 10 Mar, 2004

List of tropical cyclone events considered in this study, with simulation period. All initial conditions are at time 00z (midnight UTC).

Single-cell reconstruction of the early maternal–fetal interface in humans

Roser Vento-Tormo^{1,2,13}, Mirjana Efremova^{1,13}, Rachel A. Botting³, Margherita Y. Turco^{2,4,5}, Miquel Vento-Tormo⁶, Kerstin B. Meyer¹, Jong-Eun Park¹, Emily Stephenson³, Krzysztof Polański¹, Angela Goncalves^{1,7}, Lucy Gardner^{2,4}, Staffan Holmqvist⁸, Johan Henriksson¹, Angela Zou¹, Andrew M. Sharkey^{2,4}, Ben Millar³, Barbara Innes³, Laura Wood¹, Anna Wilbrey-Clark¹, Rebecca P. Payne³, Martin A. Ivarsson⁴, Steve Lisgo⁹, Andrew Filby³, David H. Rowitch⁸, Judith N. Bulmer³, Gavin J. Wright¹, Michael J. T. Stubbington¹, Muzlifah Haniffa^{1,3,10,14*}, Ashley Moffett^{2,4,14*} & Sarah A. Teichmann^{1,11,12,14*}

During early human pregnancy the uterine mucosa transforms into the decidua, into which the fetal placenta implants and where placental trophoblast cells intermingle and communicate with maternal cells. Trophoblast–decidual interactions underlie common diseases of pregnancy, including pre-eclampsia and stillbirth. Here we profile the transcriptomes of about 70,000 single cells from first-trimester placentas with matched maternal blood and decidual cells. The cellular composition of human decidua reveals subsets of perivascular and stromal cells that are located in distinct decidual layers. There are three major subsets of decidual natural killer cells that have distinctive immunomodulatory and chemokine profiles. We develop a repository of ligand–receptor complexes and a statistical tool to predict the cell-type specificity of cell–cell communication via these molecular interactions. Our data identify many regulatory interactions that prevent harmful innate or adaptive immune responses in this environment. Our single-cell atlas of the maternal–fetal interface reveals the cellular organization of the decidua and placenta, and the interactions that are critical for placentation and reproductive success.

During early pregnancy, the uterine mucosal lining—the endometrium—is transformed into the decidua under the influence of progesterone. Decidualization results from a complex and well-orchestrated differentiation program that involves all cellular elements of the mucosa: stromal, glandular and immune cells, the last of which include the distinctive decidual natural killer (dNK) cells^{1,2}. The blastocyst implants into the decidua, and initially—before arterial connections are established—uterine glands are the source of histotrophic nutrition in the placenta^{3,4}. After implantation, placental extravillous trophoblast cells (EVT) invade through the decidua and move towards the spiral arteries, where they destroy the smooth muscle media and transform the arteries into high conductance vessels⁵. Balanced regulation of EVT invasion is critical to pregnancy success: to ensure correct allocation of resources to mother and baby, arteries must be sufficiently transformed but excessive invasion must be prevented⁶. The pivotal regulatory role of the decidua is obvious from the life-threatening, uncontrolled trophoblast invasion that occurs when the decidua is absent, as when the placenta implants on a previous Caesarean section scar⁷.

EVT have a unique human leukocyte antigen (HLA) profile: they do not express the dominant T cell ligands, class I HLA-A and HLA-B, or class II molecules^{8,9} but do express HLA-G and HLA-E and polymorphic HLA-C class I molecules. These trophoblast HLA ligands have receptors that are expressed by the dominant decidual immune cells (that is, dNKs), including maternal killer immunoglobulin-like receptors (KIRs) some of which bind to HLA-C molecules^{10,11}. Certain combinations of maternal KIRs and fetal HLA-C genetic variants are

associated with pregnancy disorders such as pre-eclampsia, in which trophoblast invasion is deficient¹². However, detailed understanding of the cellular interactions in the decidua that support early pregnancy is lacking.

In this study, we used single-cell transcriptomics to comprehensively resolve the cell states that are involved in maternal–fetal communication in the decidua, during early pregnancy when the placenta is established. We then used a computational framework to predict cell-type-specific ligand–receptor complexes and present a new database of the curated complexes (www.CellPhoneDB.org/). By integrating these predictions with spatial *in situ* analysis, we construct a detailed molecular and cellular map of the human decidual–placental interface.

Maternal and fetal cells in early pregnancy

We combined droplet-based encapsulation (using the 10x Genomics Chromium system)¹³ and plate-based Smart-seq2¹⁴ single-cell transcriptome profiles from the maternal–fetal interface (11 decidua and 5 placentas from 6–14 gestational weeks) and six matched peripheral blood mononuclear cells (Fig. 1a, b, Supplementary Tables 1, 2, Extended Data Fig. 1). After computational quality control and integration of transcriptomes from both technologies, we performed graph-based clustering (see Methods) of the combined dataset and used cluster-specific marker genes to annotate the clusters (Fig. 1c, Extended Data Figs. 2, 3a–d, Supplementary Table 2). We studied T cell composition and clonal expansion using full-length transcriptomes from

¹Wellcome Sanger Institute, Cambridge, UK. ²Centre for Trophoblast Research, University of Cambridge, Cambridge, UK. ³Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, UK. ⁴Department of Pathology, University of Cambridge, Cambridge, UK. ⁵Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK. ⁶YDEVS software development, Valencia, Spain. ⁷German Cancer Research Center (DKFZ), Heidelberg, Germany. ⁸Department of Paediatrics, Wellcome - MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. ⁹Human Developmental Biology Resource, Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK. ¹⁰Department of Dermatology and NIHR Newcastle Biomedical Research Centre, Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. ¹¹Theory of Condensed Matter Group, The Cavendish Laboratory, University of Cambridge, Cambridge, UK. ¹²European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. ¹³These authors contributed equally: Roser Vento-Tormo, Mirjana Efremova. ¹⁴These authors jointly supervised this work: Muzlifah Haniffa, Ashley Moffett, Sarah A. Teichmann. *e-mail: m.a.haniffa@newcastle.ac.uk; am485@cam.ac.uk; st9@sanger.ac.uk

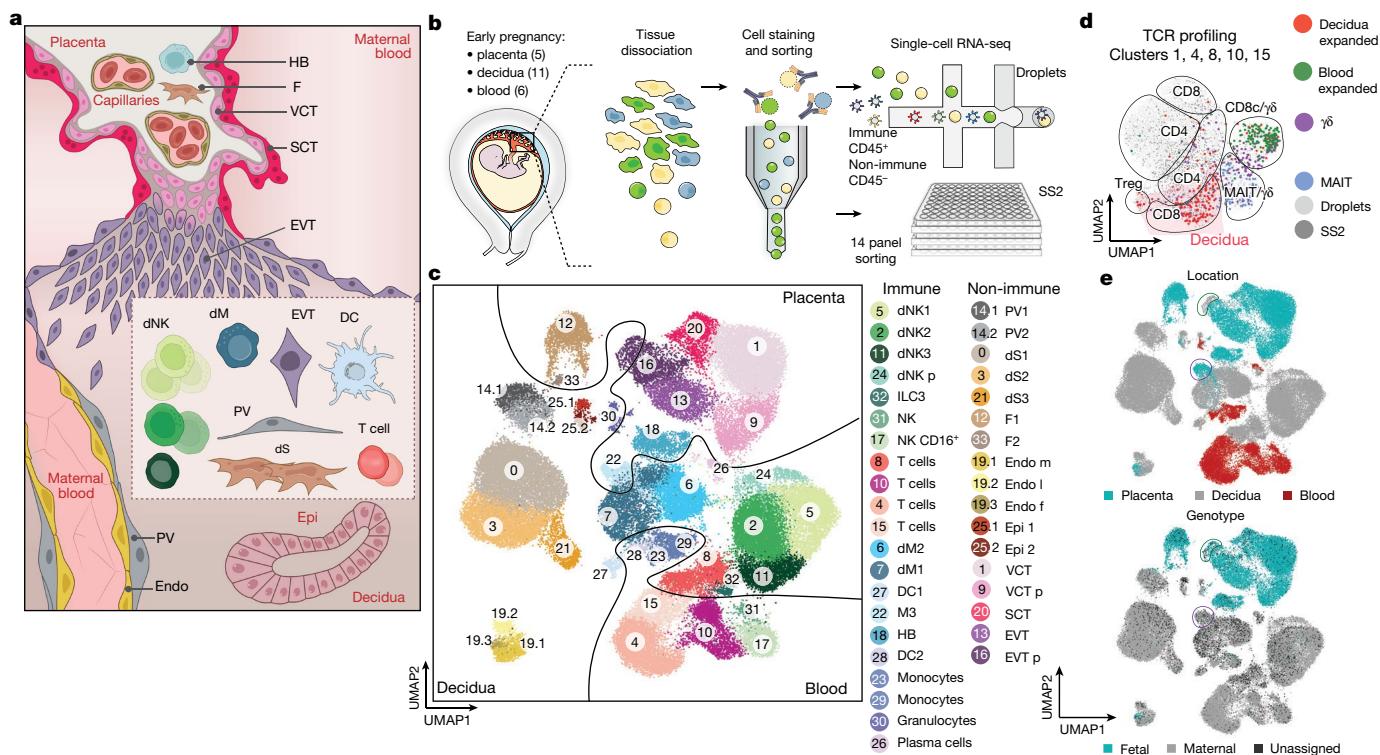


Fig. 1 | Identification of cell types at the maternal–fetal interface. **a**, Diagram illustrating the decidual–placental interface in early pregnancy. DC, dendritic cells; dM, decidual macrophages; dS, decidual stromal cells; Endo, endothelial cells; Epi, epithelial glandular cells; F, fibroblasts; HB, Hofbauer cells; PV, perivascular cells; SCT, syncytiotrophoblast; VCT, villous cytotrophoblast; EVT, extravillous trophoblast. **b**, Workflow for single-cell transcriptome profiling of decidua, placenta and maternal peripheral blood mononuclear cells. Numbers in parentheses indicate number of individuals analysed. **c**, Placental and decidual cell clusters

Smart-seq2 and reconstructed the T cell receptor sequences from this data, which showed expansion of CD8 T cells in the decidua (Fig. 1d).

We aligned single-cell RNA-seq (scRNA-seq) reads from each cell with overlapping single nucleotide polymorphisms called from maternal and fetal genomic DNA to assign cells as fetal or maternal (Fig. 1e, Extended Data Fig. 3e). As expected, decidual samples contained mostly maternal cells with a few fetal *HLA-G*⁺ EVT. Fetal cells dominate the placental samples, with the exception of maternal macrophages (M3 cluster) that express *CD14*, *S100A9*, *CD163*, *CD68* and *CSF1R* (Extended Data Fig. 3f). These are probably derived from blood monocytes incorporated into the syncytium¹⁵.

Cell communication predicted by CellPhoneDB

To systematically study the interactions between fetal and maternal cells in the decidual–placental interface, we developed a repository (www.CellPhoneDB.org) of ligand–receptor interacting pairs that accounts for their subunit architecture, representing heteromeric complexes accurately (Extended Data Fig. 4a). Both secreted and cell-surface molecules are considered; the repository therefore encompasses ligand–receptor interactions mediated by the diffusion of secreted molecules. Our repository forms the basis of a computational approach to identify biologically relevant ligand–receptor complexes. We consider the expression levels of ligands and receptors within each cell type, and use empirical shuffling to calculate which ligand–receptor pairs display significant cell-type specificity (Extended Data Fig. 4b, see Methods). This predicts molecular interactions between cell populations via specific protein complexes, and generates a potential cell–cell communication network in the decidua and placenta (Extended Data Fig. 4c–e, Supplementary Tables 3, 4).

from 10x Genomics and Smart-seq2 (SS2) scRNA-seq analysis visualized by UMAP. Colours indicate cell type or state; $n = 11$ decidua, $n = 5$ placentas and $n = 6$ blood samples. f, fetal; ILC, innate lymphocytic cells; l, lymphatic; m, maternal; p, proliferative; M3, maternal macrophages. **d**, UMAP visualization of T cell clonal expansion and clusters by integrating Smart-seq2 and 10x Genomics T cell data from clusters 4, 8, 10 and 15 from c. TCR, T cell receptor. MAIT, mucosal-associated invariant T cell. **e**, Origin of droplet cells in c by tissue (above) or genotype (below). Purple circle, maternal cells in placenta; green circle, fetal cells in decidua.

Trophoblast differentiation by scRNA-seq

To investigate maternal–fetal interactions at the decidual–placental interface, we first analysed fetal trophoblast cells isolated from placental and decidual samples: the latter contain invasive EVT (Extended Data Fig. 5a, b). Consistent with previous results^{16,17}, we resolved two distinct trophoblast differentiation pathways (Fig. 2a). As expected, decidual EVT are at the end of the trajectory, have high levels of expression of *HLA-G* and no longer express cell-cycle genes (Extended Data Fig. 5c). For villous cytotrophoblast cells, CellPhoneDB predicts interactions of receptors involved in cellular proliferation and differentiation (*EGFR*, *NRP2* and *MET*) with their corresponding ligands expressed by other cells in the placenta. *HBEGF*, potentially interacting with *EGFR*, is expressed by Hofbauer cells, and *PGF* and *HGF*—the respective ligands of *NRP2* and *MET*—are expressed by different placental fibroblast subsets (Fig. 2b, Supplementary Table 5).

By contrast, during EVT differentiation there is upregulation of receptors involved in immunomodulation, cellular adhesion and invasion, the ligands of which are expressed by decidual cells (Fig. 2b). For example, *ACKR2* is a decoy receptor for inflammatory cytokines that are produced by maternal immune cells¹⁸ and *CXCR6* is a chemokine receptor that binds to *CXCL16* expressed by the maternal macrophages. Expression of *TGFB1*—the function of which is to suppress immune responses¹⁹ and activate epithelial–mesenchymal transitions—and its receptor increases as EVT differentiate. Components involved in the epithelial–mesenchymal-transition program are upregulated at the end of the trajectory²⁰ (Extended Data Fig. 5d); these include *PAPPA* and *PAPPA2*, which encode metalloproteinases that are known to be involved in cellular invasion. In pregnancy, a decreased level of *PAPPA* is a biomarker for pre-eclampsia and fetal growth restriction, which are associated with defective EVT invasion²¹.

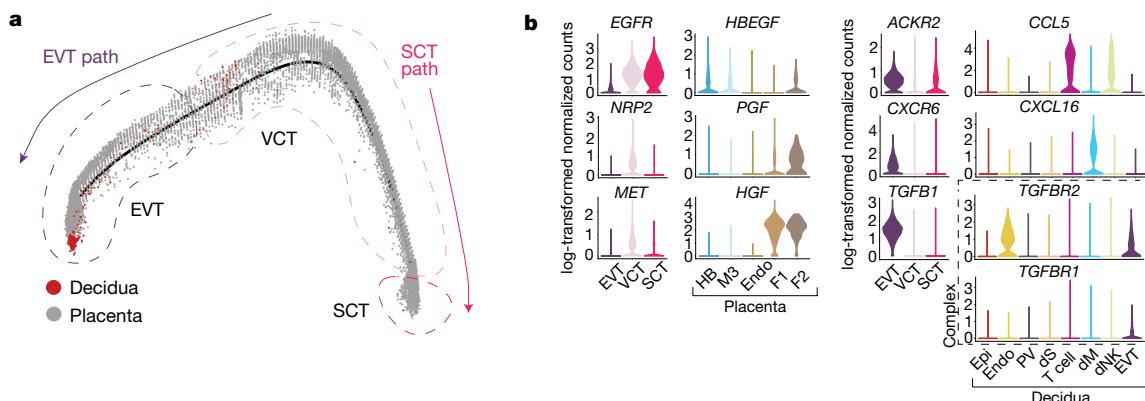


Fig. 2 | Ligand–receptor expression during EVT differentiation. **a**, Pseudotime ordering of trophoblast cells reveals EVT and SCT pathways. Enriched EPCAM⁺ and HLA-G⁺ cells on placental and decidual isolates are included. $n = 11$ deciduas and $n = 5$ placentas. **b**, Violin plots showing log-transformed, normalized expression levels for selected

ligand–receptor pairs that change during pseudotime and are predicted to be significant by CellPhoneDB (EGFR, HBEGF, NRP2, PGF, MET, HGF, ACKR2, CCL5, CXCR6, CXCL16, TGFB1, TGFBR2 and TGFBR1). Cells from Fig. 1c are used for the violin plots.

Stromal cells in the two decidual layers

EVT initially invade through the surface epithelium into the decidua compacta. Beneath this is the decidua spongiosa that contains hypersecretory glands, which provide histotrophic nutrition to the early conceptus. Markers that distinguish the different decidual fibroblast populations identify two clusters of perivascular cells (referred to as PV1 and PV2) that share expression of the smooth muscle marker (*MGP*) and are distinguished by different levels of *MCAM*, which is higher in PV1, and *MMP11*, which is higher in PV2 (Fig. 3a, Supplementary Table 6). There are three clusters of stromal cells (labelled dS1, dS2 and dS3), all of which express the WNT inhibitor *DKK1*. dS1 shares the expression of *ACTA2* and *TAGLN* with PV1 and

PV2, and lacks expression of the classical decidual markers *PRL* and *IGFBP1*. By contrast, dS2 and dS3 express *IGFBP1*, *IGFBP2* and *IGFBP6* and share markers with two subsets of decidualized stromal cells that have recently been described in vitro²². The dS3 subset expresses *PRL* as well as genes involved in steroid biosynthesis (for example, *CYP11A1*) (Extended Data Fig. 6a).

To locate the different perivascular and stromal populations in situ, we used immunohistochemistry as well as multiplexed single-molecule fluorescent in situ hybridization (smFISH) for selected markers on serial sections of decidua parietalis. These experiments confirm that cells that express *ACTA2* and *MCAM* are present in the smooth muscle media of the spiral arteries²³ and show that *MMP11* is also

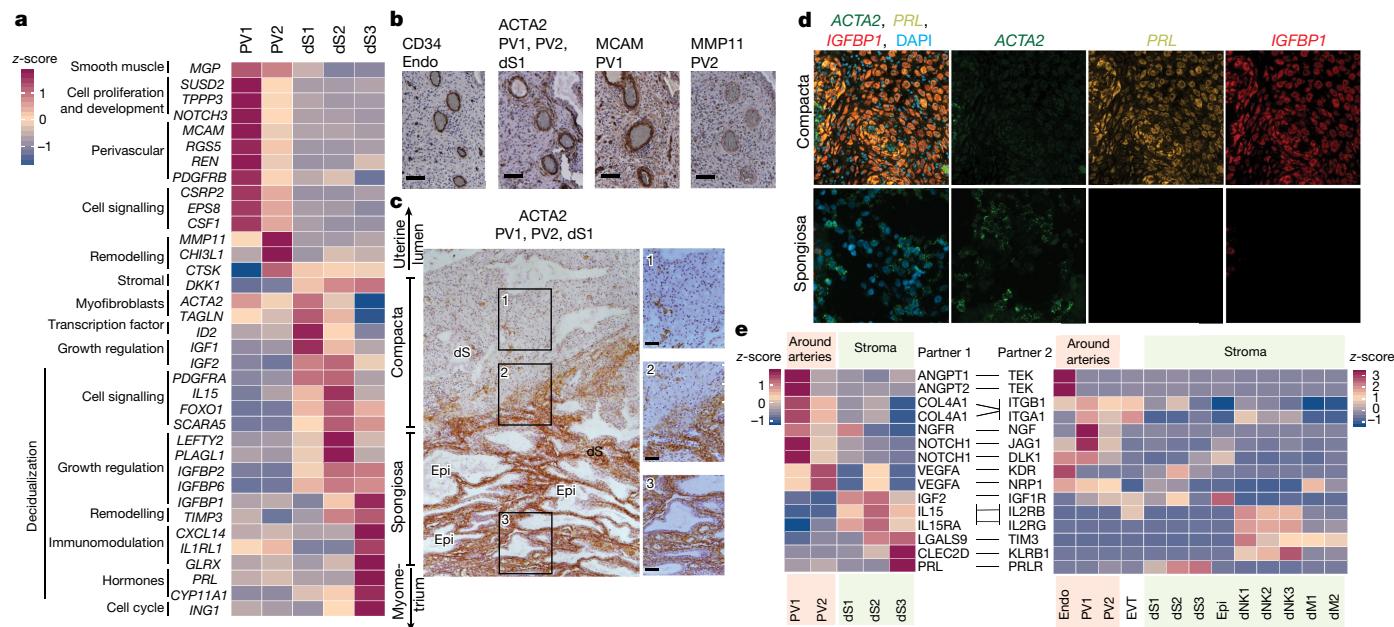


Fig. 3 | Stromal distribution in the two distinct decidual layers. **a**, Heat map showing relative expression (z-score) of selected genes for perivascular and decidual stromal cells ($n = 11$ deciduas; adjusted P value < 0.1 ; Wilcoxon rank-sum test with Bonferroni correction).

b, Immunohistochemistry of a spiral artery in serial sections of the decidua, stained for CD34 (endothelial cells), ACTA2 (PV cells and dS1 cells), MCAM (PV1 cells) and MMP11 (PV2 cells) ($n = 2$ biological replicates). Scale bar, 100 μ m. **c**, Immunohistochemistry of decidual sections stained for ACTA2, which distinguishes between ACTA2⁺ dS1 in decidua spongiosa and ACTA2[–] dS2 and dS3 in decidua compacta ($n = 3$ biological replicates). Right panels are a higher magnification of the respectively numbered inset. Scale bar, 50 μ m. **d**, Multiplexed smFISH of decidua parietalis showing two decidual layers. ACTA2⁺ dS1 in decidua spongiosa (40 \times objective); *IGFBP1*⁺ and *PRL*⁺ dS2 and dS3 confined to decidua compacta (20 \times objective) ($n = 2$ biological replicates). **e**, Heat map shows selected significant ligand–receptor interactions ($n = 6$ deciduas, P value < 0.05 , permutation test, see Methods) between PV cells and dS cells (left) and decidual cells (right) ($n = 11$ deciduas). Assays were carried out at the mRNA level, but are extrapolated to protein interactions.

biological replicates). Right panels are a higher magnification of the respectively numbered inset. Scale bar, 50 μ m. **d**, Multiplexed smFISH of decidua parietalis showing two decidual layers. ACTA2⁺ dS1 in decidua spongiosa (40 \times objective); *IGFBP1*⁺ and *PRL*⁺ dS2 and dS3 confined to decidua compacta (20 \times objective) ($n = 2$ biological replicates). **e**, Heat map shows selected significant ligand–receptor interactions ($n = 6$ deciduas, P value < 0.05 , permutation test, see Methods) between PV cells and dS cells (left) and decidual cells (right) ($n = 11$ deciduas). Assays were carried out at the mRNA level, but are extrapolated to protein interactions.

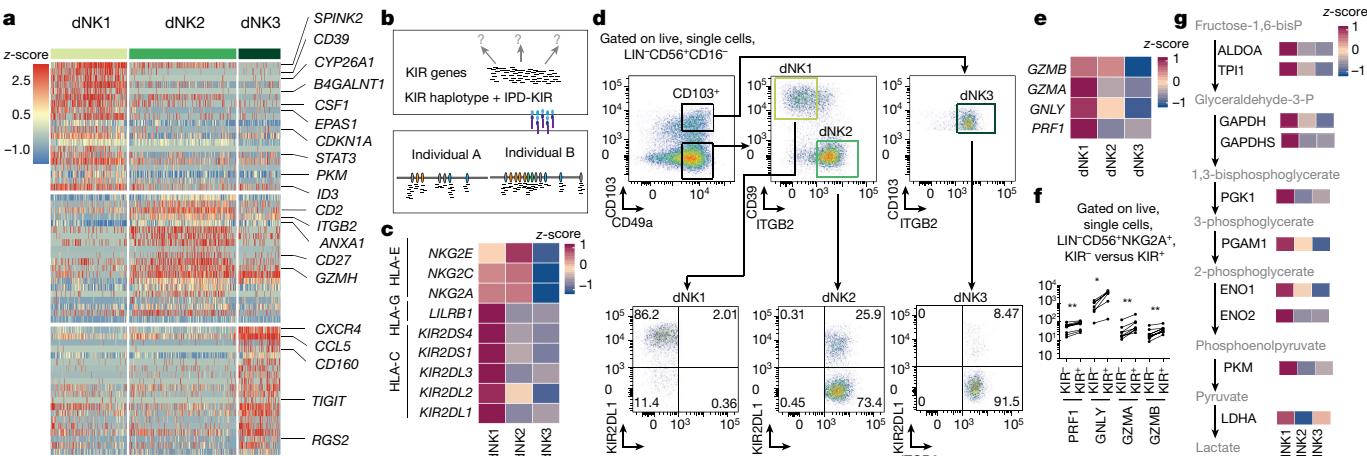


Fig. 4 | Three dNK populations. **a**, Heat map showing relative expression (z-score) of markers defining the three dNK subsets ($n = 11$ decidua; percentage 1 $> 10\%$, percentage 2 $< 60\%$; refers to the percentage of cells with expression above 0 in the corresponding cluster and all other clusters; P value < 0.1 after Bonferroni correction, Wilcoxon rank-sum test). **b**, Workflow for KIRid method (see <https://github.com/Teichlab/KIRid>). IPD-KIR, database for human KIR (available at <https://www.ebi.ac.uk/ipd/kir/>). **c**, z -scores of KIR receptors (mean expression levels). Expression values were generated using Smart-seq2 data and the KIRid approach ($n = 5$ decidua). **d**, FACS gating strategy to identify dNK subsets

present, which demonstrates that both PV1 and PV2 are perivascular (Fig. 3b). *ACTA2*⁺ dS1 cells are present between glands in the decidua spongiosa, whereas *IGFBP1*⁺ and *PRL*⁺ dS2 and dS3 cells are located in decidua compacta (Fig. 3c, d, Extended Data Fig. 7). *CYP11A1* is also expressed more abundantly in decidua compacta than in decidua spongiosa (Extended Data Fig. 6b).

Our CellPhoneDB tool predicts that the cognate receptors for angiogenic factors that are expressed by PV1 and PV2 (for example, *ANGPT1* and *VEGFA*) are located in the endothelium (Fig. 3e). EVT first invade the decidua compacta, where dS2 and dS3 express high levels of *LGALS9* and *CLEC2D*. These molecules could interact with their respective inhibitory receptors *TIM3* (also known as *HAVCR2*) and *KLRB1*—which are expressed by subsets of dNKs—enabling the stroma to suppress inflammatory reactions in the decidua.

Three decidua NK cell states

We identified three main dNK subsets (dNK1, dNK2 and dNK3), which all co-express the tissue-resident markers *CD49A* (also known as *ITGA1*) and *CD9* (Extended Data Fig. 8a). dNK1 cells express *CD39* (also known as *ENTPD1*), *CYP26A1* and *B4GALNT1*, whereas the defining markers of dNK2 cells are *ANXA1* and *ITGB2*; the latter is shared with dNK3 cells (Fig. 4a, Supplementary Table 7). dNK3 cells express *CD160*, *KLRB1* and *CD103* (also known as *ITGAE*), but not the innate lymphocyte cell marker *CD127* (also known as *IL7R*) (Extended Data Fig. 8a).

Genes of the KIR family are polymorphic and highly homologous, which makes the quantification of mRNA expression of individual KIR genes challenging¹². We therefore developed ‘KIRid’, a method that uses full-length transcript Smart-seq2 data to map the single-cell reads of each donor to the corresponding donor-specific reference of KIR alleles (Fig. 4b, see Methods). We find that dNK1 cells express higher levels of KIRs that can bind to HLA-C molecules: inhibitory *KIR2DL1*, *KIR2DL2* and *KIR2DL3* and activating *KIR2DS1* and *KIR2DS4* (Fig. 4c, Supplementary Table 8). *LILRB1*, the receptor with high affinity for the dimeric form of HLA-G molecules, is expressed only by the dNK1 subset. Both dNK1 and dNK2—but not dNK3—express activating *NKG2C* (also known as *KLRC2*) and *NKG2E* (also known as *KLRC3*) as well as inhibitory *NKG2A* (also known as *KLRC1*) receptors for HLA-E molecules (Fig. 4c). These

(representative sample from $n = 6$ individuals; Supplementary Table 9). **e**, z -scores of expression of granule molecules *PRF1*, *GNLY*, *GZMA* and *GZMB* in dNK subsets ($n = 11$ individuals). **f**, Flow cytometry to compare staining of granule components in *NKG2A*⁺*KIR*⁺ versus *NKG2A*⁺*KIR*[−] dNK cells (*PRF1* $n = 9$ individuals; *GNLY* $n = 7$ individuals; *GZMA* $n = 8$ individuals; *GZMB* $n = 10$ individuals; Supplementary Table 9). Non-parametric paired Wilcoxon test. * $P < 0.05$, ** $P < 0.01$. **g**, Right, z -scores of glycolysis enzymes (mean mRNA expression). Left, only differentially expressed enzymes are shown in the glycolysis pathway ($n = 11$ decidua; P value < 0.1 after Bonferroni correction, Wilcoxon rank-sum test).

results predict a likely function of dNK1 in the recognition and response to EVT.

To investigate these three dNK populations further, we analysed six decidua samples by flow cytometry using *CD49a* (expressed by resident dNKs), combined with markers for each dNK subset predicted from our transcriptomics data (*CD39*, *ITGB2*, *CD103* and *KIR2DL1*) (Fig. 4d, Extended Data Fig. 8b). We confirmed the presence of the three dNK populations by flow cytometry and the preferential expression of *KIR2DL1* in dNK1 (Fig. 4d, Supplementary Table 9). We analysed the morphology of dNK subsets by Giemsa staining of cells isolated by flow cytometric sorting (Extended Data Fig. 8c). dNK1 contains more cytoplasmic granules than dNK2 and dNK3, which is consistent with our scRNA-seq data that show higher levels of expression of *PRF1*, *GNLY*, *GZMA* and *GZMB* RNA in this subset (Fig. 4e). Higher levels of expression of the granule proteins (*PRF1*, *GNLY*, *GZMA* and *GZMB*) are found in *KIR*⁺ compared to *KIR*[−] dNK cells by flow cytometry (Fig. 4f). dNK1 cells also express high levels of enzymes involved in glycolysis (Fig. 4g). Thus, dNK1 cells are characterized by active glycolytic metabolism, and show higher expression of KIR genes (*KIR2DS1*, *KIR2DS4*, *KIR2DL1*, *KIR2DL2* and *KIR2DL3*), *LILRB1* and cytoplasmic granule proteins, suggesting that it is dNK1 cells that particularly interact with EVT.

First pregnancies are associated with lower proportions of dNK cells that express *LILRB1*²⁴, lower birth weights and increased occurrence of disorders such as pre-eclampsia²⁵. Metabolomic programming of mature ‘memory’ natural killer cells also occurs in chronic human cytomegalovirus infection²⁶. Together, these findings are consistent with the ‘priming’ of dNK1 cells during a first pregnancy so they can respond more effectively to the implanting placenta in subsequent pregnancies.

Immunomodulation during early pregnancy

We next used CellPhoneDB to identify the expression of cytokines and chemokines by dNKs, and to predict their interactions with other cells at the maternal–fetal interface (Fig. 5a, Extended Data Fig. 9a). However, contrary to previous studies^{24,27}, we find no evidence for substantial *VEGFA* or *IFNG* expression by dNKs *in vivo*—probably because these studies used dNK cells cultured with *IL-2* or *IL-15* *in vitro*.

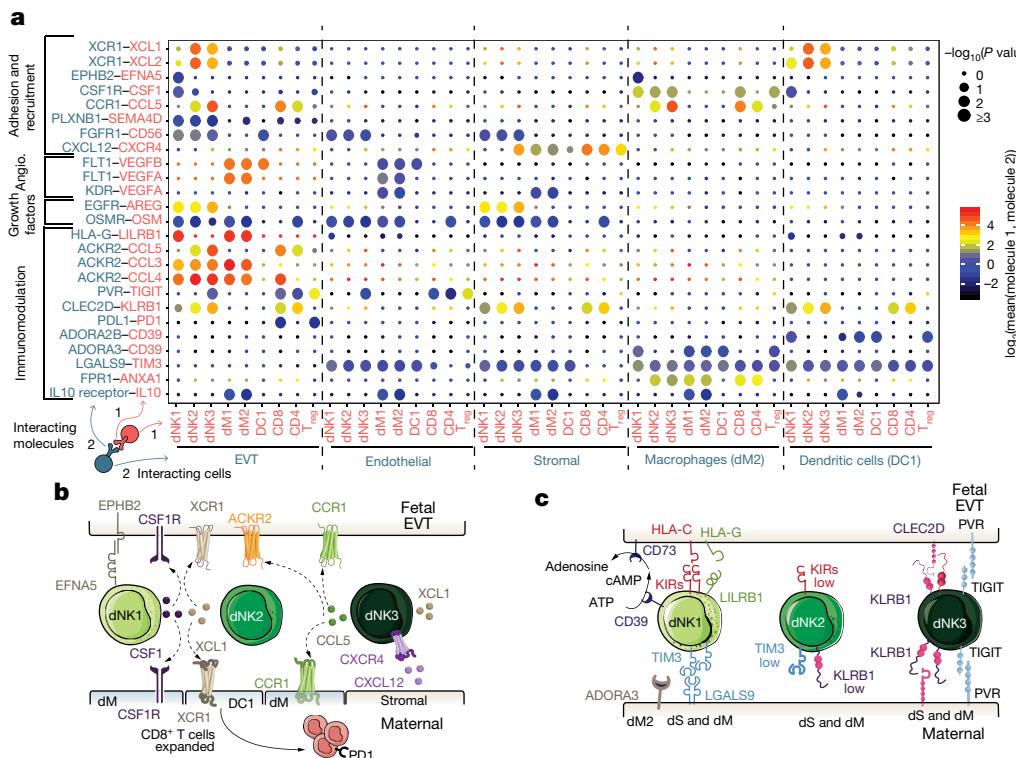


Fig. 5 | Multiple regulatory immune responses at the site of placentaion. **a**, Overview of selected ligand–receptor interactions; *P* values indicated by circle size, scale on right (permutation test, see Methods). The means of the average expression level of interacting molecule 1 in cluster 1 and interacting molecule 2 in cluster 2 are indicated by colour. Only droplet data were used ($n=6$ decidua).

dNK1 cells express higher levels of *CSF1*, the receptor of which (*CSF1R*) is expressed by EVT and macrophages (Fig. 5a, b). Secretion of *CSF1* by dNK cells and interaction with the *CSF1R* on EVT have previously been described^{28,29}, and we now pinpoint this interaction specifically to the dNK1 subset. By contrast, dNK2 and dNK3 express high levels of *XCL1*, and *CCL5* is highly expressed by dNK3 (Fig. 5a, b, Extended Data Fig. 9b). *CCR1*, the receptor for *CCL5*, is expressed by EVT, which suggests a role for dNK3 in regulating EVT invasion³⁰. The expression pattern of the *XCL1*–*XCR1* ligand–receptor complex suggests functional interactions between dNK2 and dNK3 and both EVT and conventional DC1 (labelled as DC1). DC1 recruitment, which is mediated by natural killer cells, occurs in tumour microenvironments³¹. We find an increased proportion of DC1 compared to DC2—which possibly leads to the expansion of decidual CD8⁺ T cells (Fig. 1d)—but co-expression of *PDI* (also known as *PDCL1*) suggests that local T cell activation is limited.

Our results collectively suggest that in the decidua microenvironment all damaging maternal T or natural killer cell responses to fetal trophoblast cells are prevented. There is high expression of *PDL1* (also known as *CD274*) in EVT, which we confirmed *in situ* by using immunohistochemistry on serial sections of decidua basalis (the site of trophoblast invasion) stained for *PDL1* and *HLA-G* (Extended Data Fig. 9c). We also identified putative inhibitory interactions between dNKs and EVT, in addition to the previously discussed receptor–ligand complexes between *KIR2DL1*, *KIR2DL2* or *KIR2DL3* and *HLA-C*. These include *KLRL1* and *TIGIT*, which are highly expressed by dNK3 cells, potentially binding *CLEC2D* and *PVR*, which are expressed by EVT (Fig. 5a).

We predict that the immune microenvironment of the decidua prevents inflammatory responses that could potentially be triggered by trophoblast invasion and destruction of the smooth muscle media of the spiral arteries by trophoblast (Fig. 5c). Subsets of decidual macrophages

Angio., angiogenesis. Assays were carried out at the mRNA level, but are extrapolated to protein interactions. **b**, Diagram of the main receptors and ligands expressed on the three dNK subsets that are involved in adhesion and cellular recruitment. **c**, Diagram of the main receptors and ligands expressed on the three dNK subsets that are involved in immunomodulation.

express immunomodulatory molecules such as *IL10*, the receptor of which is expressed by EVT and by maternal endothelial, stromal and myeloid cells. dNK1 cells express high levels of *SPINK2*, and dNK2 and dNK3 cells express high levels of *ANXA1*. Both of these genes encode proteins that have anti-inflammatory roles, such as inhibiting kallikreins³². The dNK1 subset expresses *CD39* (which is encoded by *ENTPD1*), which—together with *CD73* (which is encoded by *NT5E*)—converts ATP to adenosine to prevent immune activation³³ (Fig. 5c, Extended Data Fig. 9b). Expression of *CD73* is high in epithelial glands and EVT, and the adenosine receptor (*ADORA3*) is present in macrophages (Fig. 5c, Extended Data Fig. 9b). *KIR2DL1*⁺ dNK1 cells are in close physical contact with *HLA-G*⁺ EVT (Extended Data Fig. 9d), which suggests that together they could convert extracellular ATP—an inflammatory signal released upon cell death—to adenosine³⁴.

Discussion

Reproductive success depends on events that occur during placentaion in the first-trimester decidua³⁵. Other scRNA-seq studies of uterine cells in pregnancy have analysed cells at the end of gestation^{16,36} or are restricted to fetal placental populations¹⁷. To our knowledge, our study is the first comprehensive single-cell transcriptomics atlas of the maternal–fetal interface between 6–14 weeks of gestation (Extended Data Fig. 10). Similar to previous scRNA-seq analyses^{36–39}, we predict possible ligand–receptor interactions; we have developed an open repository for this purpose (www.CellPhoneDB.org/). This database accounts for the multimeric nature of ligands and receptors and is integrated with a statistical framework that predicts enriched cellular interactions between two cell types.

We show the differentiation trajectory of trophoblast cells to either villous syncytiotrophoblast (which is involved in nutrient exchange) or EVT (which invade and remodel the spiral arteries), and predict the ligand–receptor interactions that are likely to control these processes.

Our findings also suggest an environment in which any adaptive or innate immune responses that are harmful to the placenta or to the uterus are minimized. This is critical for the compromise that is needed to define the territorial boundary between mother and fetus. This environment has notable parallels with that around tumours, where inflammatory and adaptive immune responses are also dampened⁴⁰. dNK cells comprise about 70% of immune cells in the first-trimester decidua^{41,42}; we define three major subsets of dNK cells and predict that their likely function is to mediate the extent of trophoblast invasion, in addition to coordinating multiple immunomodulatory pathways that involve myeloid cells, T cells and stromal cells. Maternal immune responses are restrained by diverse classes of signalling molecules: cell-surface expression of checkpoint inhibitors such as PD1, PDL1 or TIGIT, tethered ligand–receptor complexes, secreted proteins, and small molecules such as adenosine or steroid hormones. We also show that the dNK1 subset expresses receptors for trophoblast HLA-C, HLA-E and HLA-G molecules, and can be primed metabolically through increased expression of glycolytic enzymes. The increased expression of glycolytic enzymes in dNK1 cells (which represents metabolic priming) suggests that these cells could be responsible for the different reproductive outcomes found in first compared to subsequent pregnancies.

In summary, we identify many molecular and cellular mechanisms that operate to generate a physiologically peaceful decidual environment. This cell atlas of the early maternal–fetal interface provides an essential resource for understanding normal and pathological pregnancies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0698-6>.

Received: 30 March 2018; Accepted: 15 October 2018;

Published online: 14 November 2018

1. Ramathal, C. Y., Bagchi, I. C., Taylor, R. N. & Bagchi, M. K. Endometrial decidualization: of mice and men. *Semin. Reprod. Med.* **28**, 17–26 (2010).
2. Koopman, L. A. et al. Human decidual natural killer cells are a unique NK cell subset with immunomodulatory potential. *J. Exp. Med.* **198**, 1201–1212 (2003).
3. Burton, G. J., Watson, A. L., Hempstock, J., Skepper, J. N. & Jauniaux, E. Uterine glands provide histiotrophic nutrition for the human fetus during the first trimester of pregnancy. *J. Clin. Endocrinol. Metab.* **87**, 2954–2959 (2002).
4. Hempstock, J., Cindrova-Davies, T., Jauniaux, E. & Burton, G. J. Endometrial glands as a source of nutrients, growth factors and cytokines during the first trimester of human pregnancy: a morphological and immunohistochemical study. *Reprod. Biol. Endocrinol.* **2**, 58 (2004).
5. Burton, G. J., Woods, A. W., Jauniaux, E. & Kingdom, J. C. P. Rheological and physiological consequences of conversion of the maternal spiral arteries for uteroplacental blood flow during human pregnancy. *Placenta* **30**, 473–482 (2009).
6. Fisher, S. J. Why is placentation abnormal in preeclampsia? *Am. J. Obstet. Gynecol.* **213**, S115–S122 (2015).
7. Jauniaux, E. & Burton, G. J. Placenta accreta spectrum: a need for more research on its aetiopathogenesis. *BJOG* **125**, 1449–1450 (2018).
8. Apps, R., Gardner, L. & Moffett, A. A critical look at HLA-G. *Trends Immunol.* **29**, 313–321 (2008).
9. Apps, R. et al. Human leucocyte antigen (HLA) expression of primary trophoblast cells and placental cell lines, determined using single antigen beads to characterize allotype specificities of anti-HLA antibodies. *Immunology* **127**, 26–39 (2009).
10. Sharkey, A. M. et al. Killer Ig-like receptor expression in uterine NK cells is biased toward recognition of HLA-C and alters with gestational age. *J. Immunol.* **181**, 39–46 (2008).
11. Parham, P. & Moffett, A. Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. *Nat. Rev. Immunol.* **13**, 133–144 (2013).
12. Moffett, A. & Colucci, F. Co-evolution of NK receptors and HLA ligands in humans is driven by reproduction. *Immunol. Rev.* **267**, 283–297 (2015).
13. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
14. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protocols* **9**, 171–181 (2014).
15. Burton, G. J. & Watson, A. L. The structure of the human placenta: implications for initiating and defending against virus infections. *Rev. Med. Virol.* **7**, 219–228 (1997).
16. Tsang, J. C. H. et al. Integrative single-cell and cell-free plasma RNA transcriptomics elucidates placental cellular dynamics. *Proc. Natl. Acad. Sci. USA* **114**, E7786–E7795 (2017).
17. Liu, Y. et al. Single-cell RNA-seq reveals the diversity of trophoblast subtypes and patterns of differentiation in the human placenta. *Cell Res.* **28**, 819–832 (2018).
18. Madigan, J. et al. Chemokine scavenger D6 is expressed by trophoblasts and aids the survival of mouse embryos transferred into allogeneic recipients. *J. Immunol.* **184**, 3202–3212 (2010).
19. Mariathasan, S. et al. TGF β attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature* **554**, 544–548 (2018).
20. Maltepe, E. & Fisher, S. J. Placenta: the forgotten organ. *Annu. Rev. Cell Dev. Biol.* **31**, 523–552 (2015).
21. Bolnick, J. M. et al. Altered biomarkers in trophoblast cells obtained noninvasively prior to clinical manifestation of perinatal disease. *Sci. Rep.* **6**, 32382 (2016).
22. Lucas, E. S. et al. Reconstruction of the decidual pathways in human endometrial cells using single-cell RNA-seq. Preprint at <https://www.biorxiv.org/content/early/2018/07/13/368829> (2018).
23. Muñoz-Fernández, R. et al. Human predecidual stromal cells have distinctive characteristics of pericytes: cell contractility, chemotactic activity, and expression of pericyte markers and angiogenic factors. *Placenta* **61**, 39–47 (2018).
24. Gamlie, M. et al. Trained memory of human uterine NK cells enhances their function in subsequent pregnancies. *Immunity* **48**, 951–962 (2018).
25. Kozuki, N. et al. The associations of parity and maternal age with small-for-gestational-age, preterm, and neonatal and infant mortality: a meta-analysis. *BMC Public Health* **13**, S2 (2013).
26. Cichocki, F. et al. ARID5B regulates metabolic programming in human adaptive NK cells. *J. Exp. Med.* **215**, 2379–2395 (2018).
27. Hanna, J. et al. Decidual NK cells regulate key developmental processes at the human fetal–maternal interface. *Nat. Med.* **12**, 1065–1074 (2006).
28. Jokhi, P. P., King, A., Boocock, C. & Loke, Y. W. Secretion of colony stimulating factor-1 by human first trimester placental and decidual cell populations and the effect of this cytokine on trophoblast thymidine uptake in vitro. *Hum. Reprod.* **10**, 2800–2807 (1995).
29. Hamilton, G. S., Lysiak, J. J., Watson, A. J. & Lala, P. K. Effects of colony stimulating factor-1 on human extravillous trophoblast growth and invasion. *J. Endocrinol.* **159**, 69–77 (1998).
30. Sato, Y. et al. Trophoblasts acquire a chemokine receptor, CCR1, as they differentiate towards invasive phenotype. *Development* **130**, 5519–5532 (2003).
31. Böttcher, J. P. et al. NK cells stimulate recruitment of cDC1 into the tumor microenvironment promoting cancer immune control. *Cell* **172**, 1022–1037 (2018).
32. Sotiropoulou, G. & Pampalakis, G. Kallikrein-related peptidases: bridges between immune functions and extracellular matrix degradation. *Biol. Chem.* **391**, 321–331 (2010).
33. Takenaka, M. C., Robson, S. & Quintana, F. J. Regulation of the T cell response by CD39. *Trends Immunol.* **37**, 427–439 (2016).
34. Vijayan, D., Young, A., Teng, M. W. L. & Smyth, M. J. Targeting immunosuppressive adenosine in cancer. *Nat. Rev. Cancer* **17**, 709–724 (2017).
35. Smith, G. C. S. First-trimester determination of complications of late pregnancy. *J. Am. Med. Assoc.* **303**, 561–562 (2010).
36. Pavličev, M. et al. Single-cell transcriptomics of the human placenta: inferring the cell communication network of the maternal–fetal interface. *Genome Res.* **27**, 349–361 (2017).
37. Camp, J. G. et al. Multilineage communication regulates human liver bud development from pluripotency. *Nature* **546**, 533–538 (2017).
38. Puram, S. V. et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611–1624 (2017).
39. Skelly, D. A. et al. Single-cell transcriptional profiling reveals cellular diversity and intercommunication in the mouse heart. *Cell Reports* **22**, 600–610 (2018).
40. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252–264 (2012).
41. Bulmer, J. N., Morrison, L., Longfellow, M., Ritson, A. & Pace, D. Granulated lymphocytes in human endometrium: histochemical and immunohistochemical studies. *Hum. Reprod.* **6**, 791–798 (1991).
42. King, A., Wellings, V., Gardner, L. & Loke, Y. W. Immunocytochemical characterization of the unusual large granular lymphocytes in human endometrium throughout the menstrual cycle. *Hum. Immunol.* **24**, 195–205 (1989).

Acknowledgements We thank G. Graham, J. Shilts, A. Lopez, N. Reuter, S. Orchard and P. Porras for discussions on CellPhoneDB; D. Dixon, D. Popescu, J. Fletcher, O. Chazara, L. Mamanova, A. Jinat, C. I. Mazzeo, D. McDonald and D. Bulmer for experimental help; A. Hupalowska for help with the illustrations; S. Lindsay, A. Farnworth, the HDBR, P. Ayuk and the Newcastle Uteroplacental Tissue Bank for providing samples; R. Rostom, D. McCarthy, V. Svensson, M. Hemberg and T. Gomes for computational discussions. We are indebted to the donors for participating in this research. This project was supported by ERC grants (ThDEFINE, ThSWITCH) and an EU

FET-OPEN grant (MRG-GRAMMAR no. 664918) and Wellcome Sanger core funding (no. WT206194). R.V.-T. is supported by an EMBO and HFSP Long-Term Fellowship and J.-E.P. by an EMBO Long-Term Fellowship; M.Y.T. holds a Royal Society Dorothy Hodgkin Fellowship and A.M. has a Wellcome Trust Investigator award. The human embryonic and fetal material was provided by the Joint MRC/Wellcome Trust (MR/R006237/1) HDBR.

Reviewer information *Nature* thanks B. Treutlein and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions R.V.-T. and S.A.T. conceived the study. Sample and library preparation was performed by R.V.-T. with contributions from M.Y.T., J.-E.P., E.S. and S.L.; FACS experiments were performed by R.V.-T., R.A.B., A.F., A.M.S., R.P.P. and M.A.I.; histology staining was performed by J.N.B., L.G., R.V.-T., M.Y.T., B.M., B.I., S.H., D.H.R. and A.W.-C.; M.E. and R.V.-T. analysed and interpreted the data with contributions from M.V.-T., M.J.T.S., L.W., G.J.W., A.G., A.Z., J.H., K.B.M.,

K.P., M.H., A.M. and S.A.T.; R.V.-T., A.M. and S.A.T. wrote the manuscript with contributions from M.H., M.E., K.B.M. and M.Y.T.; M.H., A.M. and S.A.T. co-directed the study. All authors read and accepted the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0698-6>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0698-6>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.H., A.M. or S.A.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Patient samples. All tissue samples used for this study were obtained with written informed consent from all participants in accordance with the guidelines in The Declaration of Helsinki 2000 from multiple centres.

Human embryo, fetal and decidual samples were obtained from the MRC and Wellcome-funded Human Developmental Biology Resource (HDBR⁴³, <http://www.hdbi.org>), with appropriate maternal written consent and approval from the Newcastle and North Tyneside NHS Health Authority Joint Ethics Committee (08/H0906/21+5). The HDBR is regulated by the UK Human Tissue Authority (HTA; www.hta.gov.uk) and operates in accordance with the relevant HTA Codes of Practice. Decidual tissue for smFISH (Extended Data Fig. 7c) was also covered by this ethics protocol.

Peripheral blood from women undergoing elective terminations was collected under appropriate maternal written consent and with approvals from the Newcastle Academic Health Partners (reference NAHPB-093) and HRA NHS Research Ethics Committee North-East-Newcastle North Tyneside 1 (REC reference 12/NE/0395)

Decidual tissue for immunohistochemistry (Fig. 3b, c, Extended Data Figs. 7a, 9c, d) and flow cytometry staining for granule proteins was obtained from elective terminations of normal pregnancies at Addenbrooke's Hospital (Cambridge) between 6 and 12 weeks gestation, under ethical approval from the Cambridge Local Research Ethics Committee (04/Q0108/23).

Decidual tissue for smFISH (Fig. 3d, Extended Data Fig. 6b, 7b) was obtained from the Newcastle Uteroplacental Tissue Bank. Ethics numbers are: Newcastle and North Tyneside Research Ethics Committee 1 Ref:10/H0906/71 and 16/NE/0167. **Isolation of decidual, placental and blood cells.** Decidual and placental tissue was washed in Ham's F12 medium, macroscopically separated and then washed for at least 10 min in RPMI or Ham's F12 medium, respectively, before processing.

Decidual tissues were chopped using scalpels into approximately 0.2-mm³ cubes and enzymatically digested in 15 ml 0.4 mg/ml collagenase V (Sigma, C-9263) solution in RPMI 1640 medium (Thermo Fisher Scientific, 21875-034)/10% FCS (Biosfera, FB-1001) at 37°C for 45 min. The supernatant was diluted with medium and passed through a 100-μm cell sieve (Corning, 431752) and then a 40-μm cell sieve (Corning, 431750). The flow-through was centrifuged and resuspended in 5 ml of red blood cell lysis buffer (Invitrogen, 00-4300) for 10 min.

Each first-trimester placenta was placed in a Petri dish and the placental villi were scraped from the chorionic membrane using a scalpel. The stripped membrane was discarded and the resultant villous tissue was enzymatically digested in 70 ml 0.2% trypsin 250 (Pan Biotech P10-025100P)/0.02% EDTA (Sigma E9884) in PBS with stirring at 37°C for 9 min. The disaggregated cell suspension was passed through sterile muslin gauze (Winware food grade) and washed through with Hams F12 medium (Biosera SM-H0096) containing 20% FBS (Biosera FB-1001). Cells were pelleted from the filtrate by centrifugation and resuspended in Ham's F12. The undigested gelatinous tissue remnant was retrieved from the gauze and further digested with 10–15 ml collagenase V at 1.0 mg/ml (Sigma C9263) in Ham's F12 medium/10% FBS with gentle shaking at 37°C for 10 min. The disaggregated cell suspension from collagenase digestion was passed through sterile muslin gauze and the cells pelleted from the filtrate as before. Cells obtained from both enzyme digests were pooled together and passed through a 100-μm cell sieve (Corning, 431752) and washed in Ham's F12. The flow-through was centrifuged and resuspended in 5 ml of red blood cell lysis buffer (Invitrogen, 00-4300) for 10 min.

Blood samples were carefully layered onto a Ficoll–Paque gradient (Amersham) and centrifuged at 2,000 r.p.m. for 30 min without breaks. Peripheral blood mononuclear cells from the interface between the plasma and the Ficoll–Paque gradient were collected and washed in ice-cold phosphate-buffered saline (PBS), followed by centrifugation at 2,000 r.p.m. for 5 min. The pellet was resuspended in 5 ml of red blood cell lysis buffer (Invitrogen, 00-4300) for 10 min.

Assignment of fetal developmental stage. Up to eight post-conception weeks, embryos are staged using the Carnegie staging method⁴⁴. At fetal stages beyond eight post-conception weeks, age was estimated from measurements of foot length and heel-to-knee length. These were compared with a standard growth chart⁴⁵.

Flow cytometry staining, cell sorting and single-cell RNA-seq. Decidual and blood cells were incubated at 4°C with 2.5 μl of antibodies in 1% FBS in DPBS without calcium and magnesium (Thermo Fisher Scientific, 14190136). DAPI was used for live versus dead discrimination. We used an antibody panel designed to enrich for certain populations for single-cell sorting and scRNA-seq. Cells were sorted using a Becton Dickinson (BD) FACS Aria Fusion with 5 excitation lasers (355 nm, 405 nm, 488 nm, 561 nm and 635 nm red), and 18 fluorescent detectors, plus forward and side scatter. The sorter was controlled using BD FACS DIVA software (version 7). The antibodies used are listed in Supplementary Table 10.

For single-cell RNA-seq using the plate-based Smart-seq2 protocol, we created overlapping gates that comprehensively and evenly sampled all immune-cell

populations in the decidua (Extended Data Fig. 1). B cells (CD19⁺ or CD20⁺) were excluded from our analysis, owing to their absence in decidua⁴⁶. Single cells were sorted into 96-well full-skirted Eppendorf plates chilled to 4°C, prepared with lysis buffer consisting of 10 μl of TCL buffer (Qiagen) supplemented with 1% β-mercaptoethanol. Single-cell lysates were sealed, vortexed, spun down at 300g at 4°C for 1 min, immediately placed on dry ice and transferred for storage at -80°C. The Smart-seq2 protocol was performed on single cells as previously described^{41,47}, with some modifications⁴⁸. Libraries were sequenced, aiming at an average depth of 1 million reads per cell, on an Illumina HiSeq 2000 with version 4 chemistry (paired-end, 75-bp reads).

For the droplet scRNA-seq methods, blood and decidual cells were sorted into immune (CD45⁺) and non-immune (CD45⁻) fractions. B cells (CD19⁺ or CD20⁺) were excluded from blood analysis, owing to their absence in decidua⁴⁶. Only viable cells were considered. Placental cells were stained for DAPI and only viable cells were sorted. To improve trophoblast trajectories, an additional enrichment of EPCAM⁺ and HLA-G⁺ was performed for selected samples (Fig. 2 only). Cells were sorted into an Eppendorf tube containing PBS with 0.04% BSA. Cells were immediately counted using a Neubauer haemocytometer and loaded in the 10x-Genomics Chromium. The 10x-Genomics v2 libraries were prepared as per the manufacturer's instructions. Libraries were sequenced, aiming at a minimum coverage of 50,000 raw reads per cell, on an Illumina HiSeq 4000 (paired-end; read 1: 26 cycles; i7 index: 8 cycles, i5 index: 0 cycles; read 2: 98 cycles).

Flow cytometry staining for granule proteins. For intracellular staining of granule proteins, dNKs were surface-stained for 30 min in FACS buffer with antibodies (listed in Supplementary Table 10). Cells were washed with FACS buffer followed by staining with dead cell marker (DCM Aqua) and streptavidin Qdot605. dNKs were then treated with FIX & PERM (Thermo Fisher Scientific) and stained for granule proteins. Samples were run on an LSRFortessa FACS analyser (BD Biosciences) and data analysed using FlowJo (Tree Star). dNKs were gated as CD3⁻CD14⁻CD19⁻ live cells; CD56⁺NKG2A⁺ and then KIR⁺ and KIR⁻ subsets were generated using Boolean functions with the gates for all the different KIRs stained (KIR⁺), and their inverse gates (KIR⁻). Wilcoxon test was used to compare granule protein staining between paired dNK subsets from the same donor. A P value < 0.05 was considered to be statistically significant.

Immunohistochemistry. Four-micrometre tissue sections from formalin-fixed, paraffin-wax-embedded human decidual and placental tissues were dewaxed with Histoclear, cleared in 100% ethanol and rehydrated through gradients of ethanol to PBS. Sections were blocked with 2% serum (of species in which the secondary antibody was made) in PBS, incubated with primary antibody overnight at 4°C and slides were washed in PBS. Biotinylated horse anti-mouse or goat anti-rabbit secondary antibodies were used, followed by Vectastain ABC-HRP reagent (Vector, PK-6100) and developed with di-aminobenzidine (DAB) substrate (Sigma, D4168). Sections were counterstained with Carazzi's haematoxylin and mounted in glycerol and gelatin mounting medium (Sigma, GG1-10). Primary antibody was replaced with equivalent concentrations of mouse or rabbit IgG for negative controls. See Supplementary Table 10 for antibody information. Tissue sections were imaged using a Zeiss Axiovert Z1 microscope and Axiovision imaging software SE64 version 4.8.

smFISH. Samples were fixed in 10% NBF, dehydrated through an ethanol series and embedded in paraffin wax. Five-millimetre samples were cut, baked at 60°C for 1 h and processed using standard pre-treatment conditions, as per the RNAscope multiplex fluorescent reagent kit version 2 assay protocol (manual) or the RNAscope 2.5 LS fluorescent multiplex assay (automated). TSA-plus fluorescein, Cy3 and Cy5 fluorophores were used at 1:1,500 dilution for the manual assay or 1:300 dilution for the automated assay. Slides were imaged on different microscopes: Hamamatsu Nanozoomer S60 (Extended Data Fig. 7c). Zeiss Cell Discoverer 7 (Fig. 4d, Extended Data Figs. 6, 7c). Filter details were as follows. DAPI: excitation 370–400, BS 394, emission 460–500; FITC: excitation 450–488, BS 490, emission 500–55; Cy3: excitation 540–570, BS 573, emission 540–570; Cy5: excitation 615–648, BS 691, emission 662–756. The camera used was a Hamamatsu ORCA-Flash4.0 V3 sCMOS camera.

Whole-genome sequencing. Tissue DNA and RNA were extracted from fresh-frozen samples using the AllPrep DNA/RNA/miRNA kit (Qiagen), following the manufacturer's instructions. Short insert (500-bp) genomic libraries were constructed, flowcells were prepared and 150-bp paired-end sequencing clusters generated on the Illumina HiSeq X platform, according to Illumina no-PCR library protocols, to an average of 30× coverage. Genotype information is provided in Supplementary Table 1.

Single cell RNA-seq data analysis. Droplet-based sequencing data were aligned and quantified using the Cell Ranger Single-Cell Software Suite (version 2.0, 10x Genomics)¹³ against the GRCh38 human reference genome provided by Cell Ranger. Cells with fewer than 500 detected genes and for which the total mitochondrial gene expression exceeded 20% were removed. Mitochondrial genes and genes that were expressed in fewer than three cells were also removed.

SmartSeq2 sequencing data were aligned with HISAT2⁴⁹, using the same genome reference and annotation as the 10x Genomics data. Gene-specific read counts were calculated using HTSeq-count⁵⁰. Cells with fewer than 1,000 detected genes and more than 20% mitochondrial gene expression content were removed. Furthermore, mitochondrial genes and genes expressed in fewer than three cells were also removed. To remove batch effects due to background contamination of cell free RNA, we also removed a set of genes that had a tendency to be expressed in ambient RNA (*PAEP*, *HBG1*, *HBA1*, *HBA2*, *HBM*, *AHSP* and *HBG2*).

Downstream analyses—such as normalization, shared nearest neighbour graph-based clustering, differential expression analysis and visualization—were performed using the R package Seurat⁵¹ (version 2.3.3). Droplet-based and SmartSeq2 data were integrated using canonical correlation analysis, implemented in the Seurat alignment workflow⁵². Cells, the expression profile of which could not be well-explained by low-dimensional canonical correlation analysis compared to low-dimensional principal component analysis, were discarded, as recommended by the Seurat alignment tutorial. Clusters were identified using the community identification algorithm as implemented in the Seurat 'FindClusters' function. The shared nearest neighbour graph was constructed using between 5 and 40 canonical correlation vectors as determined by the dataset variability; the resolution parameter to find the resulting number of clusters was tuned so that it produced a number of clusters large enough to capture most of the biological variability. UMAP analysis was performed using the RunUMAP function with default parameters. Differential expression analysis was performed based on the Wilcoxon rank-sum test. The *P* values were adjusted for multiple testing using the Bonferroni correction. Clusters were annotated using canonical cell-type markers. Two clusters of peripheral blood monocytes represented the same cell type and were therefore merged.

We further removed contaminating cells: (i) maternal stromal cells that were gathered in the placenta for one of the fetuses; (ii) a shared decidual–placental cluster with fetal cells mainly present in two fetuses (which we think is likely to be contaminating cells from other fetal tissues due to the surgical procedure). This can occur owing to the source of the tissue and the trauma of surgery. We also removed a cluster for which the top markers were genes associated with dissociation-induced effects⁵³. Each of the remaining clusters contained cells from multiple different fetuses, indicating that the cell types and states we observed are not affected by batch effects.

We found further diversity within the T cell clusters, as well as the clusters of endothelial, epithelial and perivascular cells, which we then reanalysed and partitioned separately, using the same alignment and clustering procedure.

The trophoblast clusters (clusters 1, 9, 20, 13 and 16 from Fig. 1d) were taken from the initial analysis of all cells and merged with the enriched EPCAM⁺ and HLA-G⁺ cells. The droplet-based and Smart-seq2 datasets were integrated and clustered using the same workflow as described above. Only cells that were identified as trophoblast were considered for trajectory analysis.

Trajectory modelling and pseudotemporal ordering of cells was performed with the monocle 2 R package⁵⁴ (version 2.8.0). The most highly variable genes were used for ordering the cells. To account for the cell-cycle heterogeneity in the trophoblast subpopulations, we performed hierarchical clustering of the highly variable genes and removed the set of genes that cluster with known cell-cycle genes such as *CDK1*. Genes which changed along the identified trajectory were identified by performing a likelihood ratio test using the function differentialGeneTest in the monocle 2 package.

Network visualization was done using Cytoscape (version 3.5.1). The decidua network was created considering only edges with more than 30 interactions. The networks layout was set to force-directed layout.

KIR typing. Polymerase chain reaction sequence-specific primer was performed to amplify the genomic DNA for presence or absence of 12 KIR genes (*KIR2DL1*, *KIR2DL2*, *KIR2DL3*, *KIR2DL5* (both *KIR2DL5A* and *KIR2DL5B*), *KIR3DL1*, *KIR2DS1*, *KIR2DS2*, *KIR2DS3*, *KIR2DS4*, *KIR2DS5* and *KIR3DS1*) and the pseudogene *KIR2DP1*. *KIR2DS4* alleles were also typed as being either full-length or having the 22-bp deletion that prevents cell-surface expression. Two pairs of primers were used for each gene, selected to give relatively short amplicons of 100–800 bp, as previously described⁵⁵. Extra KIR primers were designed using sequence information from the IPD-KIR database (release 2.4.0) to detect rare alleles of *KIR2DS5* and *KIR2DL3* (*KIR2DS5*, 2DS5rev2: TCC AGA GGG TCA CTG GGA and *KIR2DL3*, 2DL3rev3: AGA CTC TTG GTC CAT TAC CG)⁵⁶. KIR haplotypes were defined by matrix subtraction of gene copy numbers using previously characterized common and contracted KIR haplotypes using the KIR Haplotype Identifier software (www.bioinformatics.cimr.cam.ac.uk/haplotypes).

Inferring maternal or fetal origin of single cells from droplet-based scRNA-seq using whole-genome sequencing variant calls. To match the processing of the whole-genome sequencing datasets, droplet-based sequencing data from decidua and placenta samples were realigned and quantified against the GRCh37 human reference genome using the Cell Ranger Single-Cell Software Suite (version 2.0)¹³. The fetal or maternal origin of each barcoded cell was then determined using the

tool demuxlet⁵⁷. In brief, demuxlet can be used to deconvolve droplet-based scRNA-seq experiments in which cells are pooled from multiple genetically distinct individuals. Given a set of genotypes corresponding to these individuals, demuxlet infers the most likely genetic identity of each droplet by estimating the likelihood of observing scRNA-seq reads from the droplet overlapping known single nucleotide polymorphisms. Demuxlet inferred the identities of cells in this study by analysing each Cell Ranger-aligned BAM file from decidua and placenta in conjunction with a VCF file, containing the high-quality whole-genome-sequence variant calls from the corresponding mother and fetus. Each droplet was assigned to be maternal, fetal or unknown in origin (ambiguous or a potential doublet), and these identities were then linked with the transcriptome-based cell clustering data to confirm the maternal and fetal identity of each annotated cell type.

T cell receptor analysis by TraCeR. The T cell receptor sequences for each single T cell were assembled using TraCeR⁵⁸, which allowed the reconstruction of the T cell receptors from scRNA-seq data and their expression abundance (transcripts per million), as well as identification of the size, diversity and lineage relation of clonal subpopulations. In total, we obtained the T cell receptor sequences for 1,482 T cells with at least one paired productive $\alpha\beta$ or $\gamma\delta$ chain. Cells for which more than two recombinants were identified for a particular locus were excluded from further analysis.

Whole-genome sequencing alignment and variant calling. Maternal and fetal whole-genome sequencing data were mapped to the GRCh37.p13 reference genome using BWA-MEM version 0.7.15⁵⁹. The SAMtools⁶⁰ fixmate utility (version 1.5) was used to update read-pairing information and mate-related flags. Reads near known indels from the Mills⁶¹ and 1000G⁶² gold standard reference set for hg19/GRCh37 were locally realigned using GATK IndelRealigner version 3.7⁶³. Base-calling assessment and base-quality scores were adjusted with GATK BaseRecalibrator and PrintReads version 3.7^{60,63}. PCR duplicates were identified and removed using Picard MarkDuplicates version 2.14.1^{63,64}. Finally, bcftools mpileup and call version 1.6⁶⁵ were used to produce genotype likelihoods and output called variants at all known biallelic single nucleotide polymorphism sites that overlap protein-coding genes. For each sample, variants called with phred-scale quality score ≥ 200 , at least 20 supporting reads and mapping quality ≥ 60 were retained as high-quality variants.

Quantification of KIR gene expression by KIRid. The KIR locus is highly polymorphic in terms of both numbers of genes and alleles¹¹. Including a single reference sequence for each gene can lead to reference bias for donors that happen to better match the reference sequence. To address these issues, we used a tailored approach in which we first built a total cDNA reference by concatenating the Ensembl coding and non-coding transcript sequences, excluding transcripts belonging to the KIR genes (GRCh38, version 90), and the full set of known KIR cDNAs sequences from the IPD-KIR database⁶⁶ (release 2.7.0). For each donor, we removed transcript sequences for KIR genes determined to be absent in that individual, which decreases the extent of multi-mapping and quantification. The single-cell reads of each donor were then mapped to the corresponding donor-specific reference using Kallisto⁶⁷ (version 0.43.0 with default options). Expression levels were quantified using the multi-mapping deconvolution tool MMSEQ⁶⁸, and gene-level estimates were obtained by aggregating over different alleles for each KIR gene.

Cell-cell communication analysis. To enable a systematic analysis of cell-cell communication molecules, we developed CellPhoneDB, a public repository of ligands, receptors and their interactions. Our repository relies on the use of public resources to annotate receptors and ligands. We include subunit architecture for both ligands and receptors, to accurately represent heteromeric complexes.

Ligand–receptor pairs are defined based on physical protein–protein interactions (see sections of 'CellPhoneDB annotations'). We provide CellPhoneDB with a user-friendly web interface at www.CellPhoneDB.org, where the user can search for ligand–receptor complexes and interrogate their own single-cell transcriptomics data.

To assess cellular crosstalk between different cell types, we used our repository in a statistical framework for inferring cell-cell communication networks from single-cell transcriptome data. We derived enriched receptor–ligand interactions between two cell types based on expression of a receptor by one cell type and a ligand by another cell type, using the droplet-based data. To identify the most relevant interactions between cell types, we looked for the cell-type specific interactions between ligands and receptors. Only receptors and ligands expressed in more than 10% of the cells in the specific cluster were considered.

We performed pairwise comparisons between all cell types. First, we randomly permuted the cluster labels of all cells 1,000 times and determined the mean of the average receptor expression level of a cluster and the average ligand expression level of the interacting cluster. For each receptor–ligand pair in each pairwise comparison between two cell types, this generated a null distribution. By calculating the proportion of the means which are 'as or more extreme' than the actual mean, we obtained a *P* value for the likelihood of cell-type specificity of a given receptor–ligand

complex. We then prioritized interactions that are highly enriched between cell types based on the number of significant pairs, and manually selected biologically relevant ones. For the multi-subunit heteromeric complexes, we required that all subunits of the complex are expressed (using a threshold of 10%), and therefore we used the member of the complex with the minimum average expression to perform the random shuffling.

CellPhoneDB annotations of membrane, secreted and peripheral proteins.

Secreted proteins were downloaded from Uniprot using KW-0964 (secreted). Secreted proteins were annotated as cytokines (KW-0202), hormones (KW-0372), growth factors (KW-0339) and immune-related using Uniprot keywords and manual annotation. Cytokines, hormones, growth factors and other immune-related proteins were annotated as 'secreted highlight' proteins in our lists.

Plasma membrane proteins were downloaded from Uniprot using KW-1003 (cell membrane). Peripheral proteins from the plasma membrane were annotated using the Uniprot Keyword SL-9903, and the remaining proteins were annotated as transmembrane proteins. We completed our lists of plasma transmembrane proteins by doing an extensive manual curation using literature mining and Uniprot description of proteins with transmembrane and immunoglobulin-like domains.

Plasma membrane proteins were annotated as receptors and transporters. Transporters were defined by the Uniprot keyword KW-0813. Receptors were defined by the Uniprot keyword KW-0675. The list of receptors was extensively reviewed and new receptors were added based on Uniprot description and bibliography revision. Receptors involved in immune-cell communication were carefully annotated.

Protein lists are available at <https://www.cellphonedb.org/downloads>. Three columns indicate whether the protein has been manually curated: 'tags', 'tags_description', 'tags_reason'.

The tags column is related to the manual curation of a protein, and contains three options: (i) 'N/A', which indicates that the protein has not been manually curated; (ii) 'To_add', which indicates that secreted and/or plasma membrane protein annotation has been added; and (iii) 'To_comment', which indicates that the protein is either secreted (KW-0964) or membrane-associated (KW-1003) but that we manually added a specific property of the protein (that is, the protein is annotated as a receptor).

tags_reason is related to the protein properties, and contains five options: (i) 'extracellular_add', which indicates that the protein is manually annotated as plasma membrane; (ii) 'peripheral_add', which indicates that the protein is manually annotated as a peripheral protein instead of plasma membrane; (iii) 'secreted_add', which indicates that the protein is manually annotated as secreted; (iv) 'secreted_high', which indicates that the protein is manually annotated as secreted highlight. For cytokines, hormones, growth factors and other immune-related proteins; option (v) 'receptor_add' indicates that the protein is manually annotated as a receptor.

tags_description is a brief description of the protein, function or property related to the manually curated protein.

CellPhoneDB annotations of heteromeric receptors and ligands. Heteromeric receptors and ligands (that is, proteins that are complexes of multiple gene products) were annotated by reviewing the literature and Uniprot descriptions. Cytokine complexes, TGF family complexes and integrin complexes were carefully annotated.

If heteromers are defined in the RCSB Protein Data Bank (<http://www.rcsb.org/>), structural information is included in our CellPhoneDB annotation. Heteromeric complex lists are available at www.CellPhoneDB.org.

CellPhoneDB annotations of interactions. The majority of ligand–receptor interactions were manually curated by reviewing Uniprot descriptions and PubMed information on membrane receptors. Cytokine and chemokine interactions are annotated following the International Union of Pharmacology annotation⁶⁹. Other groups of cell-surface proteins the interactions of which were manually reviewed include the TGF family, integrins, lymphocyte receptors, semaphorins, ephrins, Notch and TNF receptors.

In addition, we considered interacting partners as: (i) binary interactions annotated by IUPHAR (<http://www.guidetopharmacology.org/>) and (ii) cytokines, hormones and growth factors interacting with receptors annotated by the iMEX consortium (<https://www.imexconsortium.org/>)⁷⁰.

We excluded from our analysis transporters and a curated list of proteins including: (i) co-receptors; (ii) nerve-specific receptors such as those related to ear-binding, olfactory receptors, taste receptors and salivary receptors, (iii) small molecule receptors, (iv) immunoglobulin chains, (v) pseudogenes and (vi) viral and retroviral proteins, pseudogenes, cancer antigens and photoreceptors. These proteins are annotated as 'others' in the protein list. We also excluded from our analysis a list of interacting partners not directly involved in cell–cell communication. The 'remove_interactions' list is available in <https://www.cellphonedb.org/downloads>.

Lists of interacting protein chains are available from <https://www.cellphonedb.org/downloads>. The column labelled 'source' indicates the curation source. Manually curated interactions are annotated as 'curated', and the bibliography used to annotate the interaction is stored in 'comments_interaction'. 'Uniprot' indicates that the interaction has been annotated using UniProt descriptions.

Linking Ensembl and Uniprot identification. We assigned to the custom-curated interaction list all the Ensembl gene identifications by matching information from Uniprot and Ensembl by the gene name.

Database structure. Information is stored in a PostgreSQL relational database (www.postgresql.org). SQLAlchemy (www.sqlalchemy.org) and Python 3 were used to build the database structure and the query logic. All the code is open source and uploaded to the webserver.

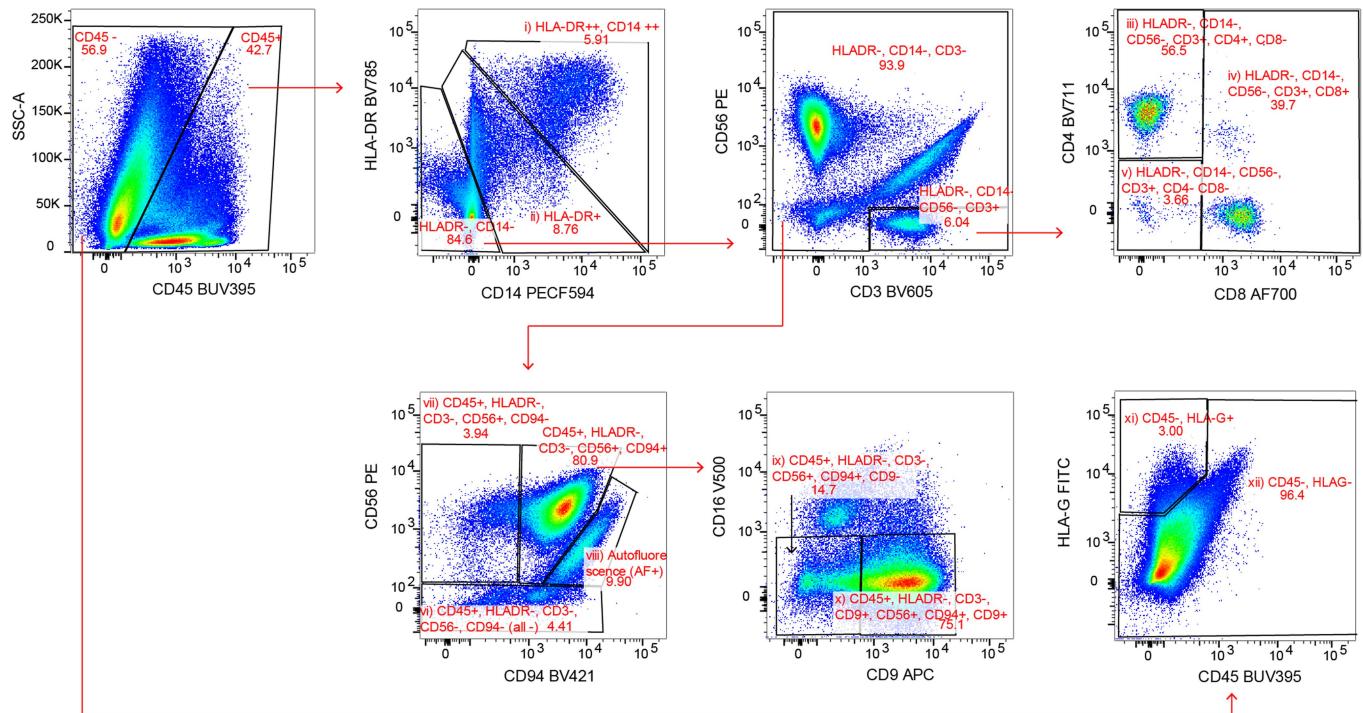
Code availability. CellPhoneDB code is available in <https://github.com/Teichlab/cellphonedb>. The code can also be downloaded from <https://cellphonedb.org/>. KIRid can be downloaded from <https://github.com/Teichlab/KIRid>.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

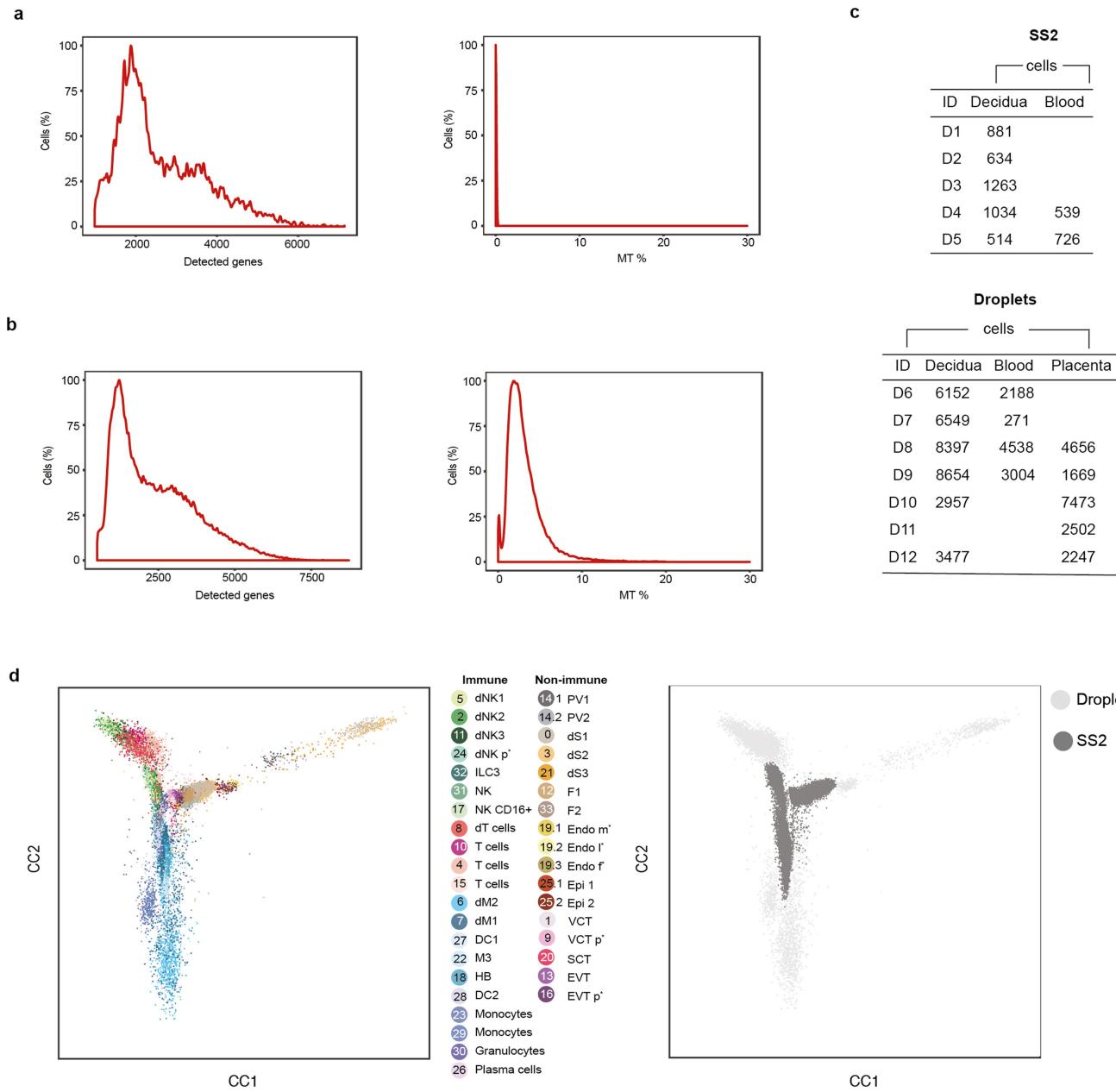
Our expression data for different tissues are also available for user-friendly interactive browsing online at <http://data.teichlab.org> (maternal–fetal interface). The raw sequencing data, expression-count data with cell classifications and the whole-genome sequencing data are deposited at ArrayExpress, with experiment codes E-MTAB-6701 (for droplet-based data), E-MTAB-6678 (for Smart-seq2 data) and E-MTAB-7304 (for the whole-genome sequencing data). Our CellPhoneDB repository is available at www.CellPhoneDB.org.

43. Gerrelli, D., Lisgo, S., Copp, A. J. & Lindsay, S. Enabling research with human embryonic and fetal tissue resources. *Development* **142**, 3073–3076 (2015).
44. O’Rahilla, R. & Muller, F. *Human Embryology and Teratology* (Wiley-Liss, New York, 1992).
45. Hern, W. M. Correlation of fetal age and measurements between 10 and 26 weeks of gestation. *Obstet. Gynecol.* **63**, 26–32 (1984).
46. Bulmer, J. N., Williams, P. J. & Lash, G. E. Immune cells in the placental bed. *Int. J. Dev. Biol.* **54**, 281–294 (2010).
47. Trombetta, J. J. et al. Preparation of single-cell RNA-seq libraries for next generation sequencing. *Curr. Protoc. Mol. Biol.* **107**, 4.22.1–4.22.17 (2014).
48. Villani, A.-C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).
49. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
50. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
51. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
52. Butler, A. & Satija, R. Integrated analysis of single cell transcriptomic data across conditions, technologies, and species. Preprint at <https://www.biorxiv.org/content/early/2017/07/18/164889> (2017).
53. van den Brink, S. C. et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).
54. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
55. Hiby, S. E. et al. Combinations of maternal KIR and fetal HLA-C genes influence the risk of preeclampsia and reproductive success. *J. Exp. Med.* **200**, 957–965 (2004).
56. Robinson, J. et al. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* **43**, D423–D431 (2015).
57. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
58. Stubbington, M. J. T. et al. T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* **13**, 329–332 (2016).
59. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <https://arxiv.org/abs/1303.3997> (2013).
60. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
61. Mills, R. E. et al. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.* **21**, 830–839 (2011).
62. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
63. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
64. Broad Institute. *Picard tools* <https://broadinstitute.github.io/picard/> (Broad Institute, 2018).
65. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
66. Robinson, J., Mistry, K., McWilliam, H., Lopez, R. & Marsh, S. G. E. IPD—the Immuno Polymorphism Database. *Nucleic Acids Res.* **38**, D863–D869 (2010).
67. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
68. Turro, E. et al. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol.* **12**, R13 (2011).
69. Bachelerie, F. et al. International Union of Basic and Clinical Pharmacology, LXXXIX. Update on the extended family of chemokine receptors and introducing a new nomenclature for atypical chemokine receptors. *Pharmacol. Rev.* **66**, 1–79 (2013).
70. Orchard, S. et al. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods* **9**, 345–350 (2012).

a**Extended Data Fig. 1 | Gating strategy for Smart-seq2 data.**

a, Gating strategy for a panel of 14 antibodies to analyse immune cells in decidua samples by Smart-seq2 (CD3, CD4, CD8, CD9, CD14, CD16, CD19, CD20, CD34, CD45, CD56, CD94, DAPI, HLA-DR and HLA-G). Cells isolated for Smart-seq2 data were gated on live; CD19- and CD20-negative, singlets and the following cell types were sorted: (i) CD45⁺CD14^{high}HLA-DR^{high}; (ii) CD45⁺HLA-DR⁺; (iii) CD45⁺HLA-DR⁻CD56⁻CD3⁺CD4⁺CD8⁻; (iv) CD45⁺HLA-DR⁻CD56⁻CD3⁺CD8⁺; (v) CD45⁺HLA-DR⁻CD56⁻CD3⁺CD4⁻CD8⁻; (vi) CD45⁺HLA-DR⁻

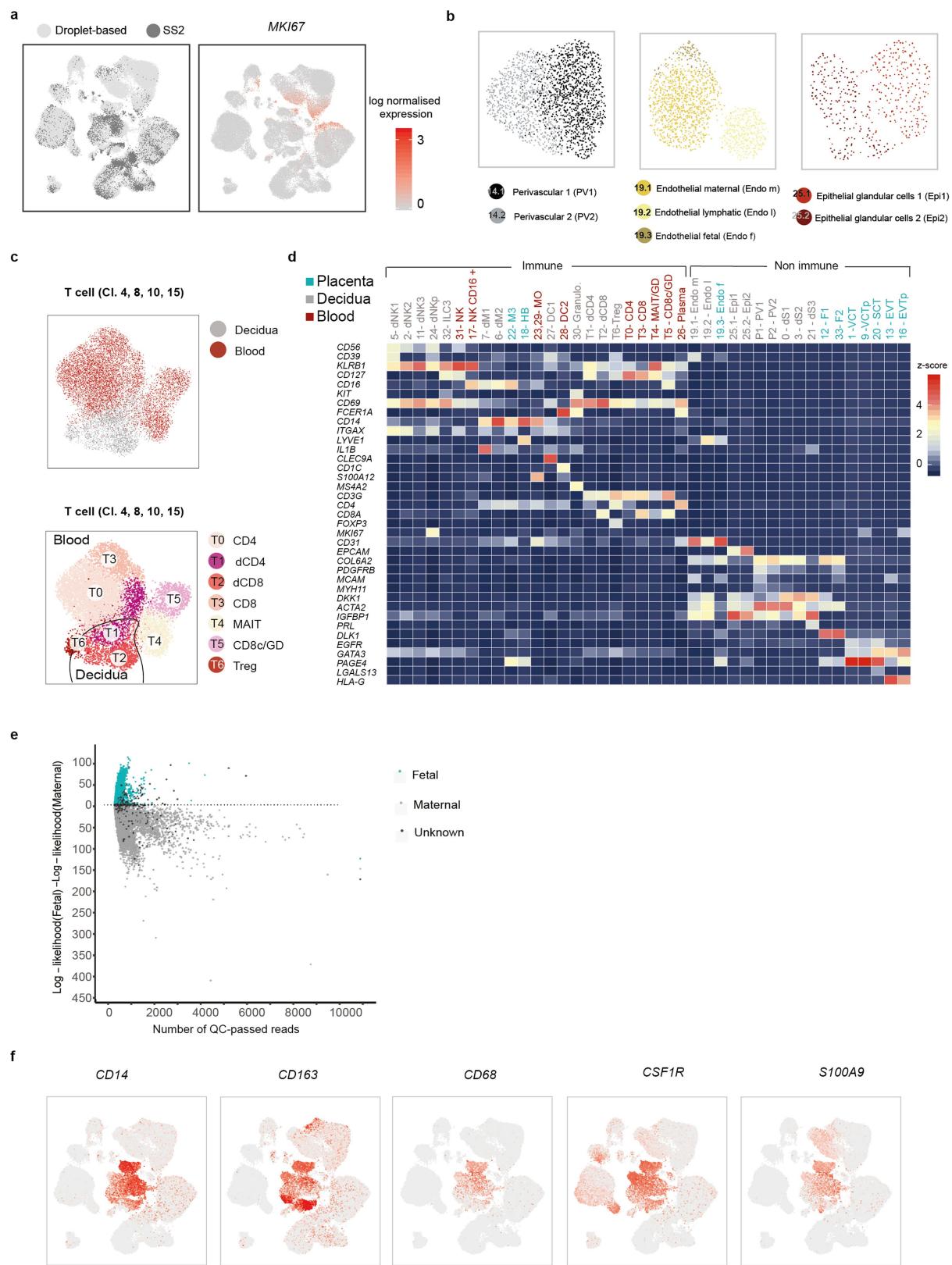
CD3⁻CD56⁻CD94⁻ (labelled 'all -' on the figure); (vii) CD45⁺HLA-DR⁻CD3⁻CD56⁺CD94⁻; (viii) autofluorescence; (ix) CD45⁺HLA-DR⁻CD3⁻CD56⁺CD94⁺CD9⁻; (x) CD45⁺HLA-DR⁻CD3⁻CD56⁺CD94⁺CD9⁺; (xi) CD45⁻HLA-G⁺; (xii) CD45⁻HLA-G⁻. Sample F9 is shown as an example. Cells from different gates were sorted in different plates: myeloid cells (gates (i) and (ii)); T cells (gates (iii), (iv) and (v)); natural killer cells (gates (vi), (vii), (viii), (ix) and (x)); CD45⁻ (gates (xi) and (xii)). Antibody information is provided in Supplementary Table 10.



* p = proliferative; m = maternal; l = lymphatic; f = fetal

Extended Data Fig. 2 | Quality control of droplet and Smart-seq2 datasets. **a**, Histograms show the distribution of the cells from the Smart-seq2 dataset ordered by number of detected genes and mitochondrial gene expression content. **b**, Histograms show the distribution of the cells from the droplet-based dataset ordered by number of detected genes and mitochondrial gene expression content. **c**, Total numbers of cells that passed the quality control, processed by Smart-seq2 and droplet scRNA-

seq. Each row is a separate donor. **d**, Canonical correlation vectors (CC1 and CC2) of integrated analysis of decidual and placental cells from the Smart-seq2 ($n = 5$ decidua, $n = 2$ peripheral blood samples) and droplet-based datasets ($n = 5$ placentas, $n = 6$ decidua and $n = 4$ blood samples), coloured on the basis of their assignment to clusters and the technology that was used for scRNA-seq.

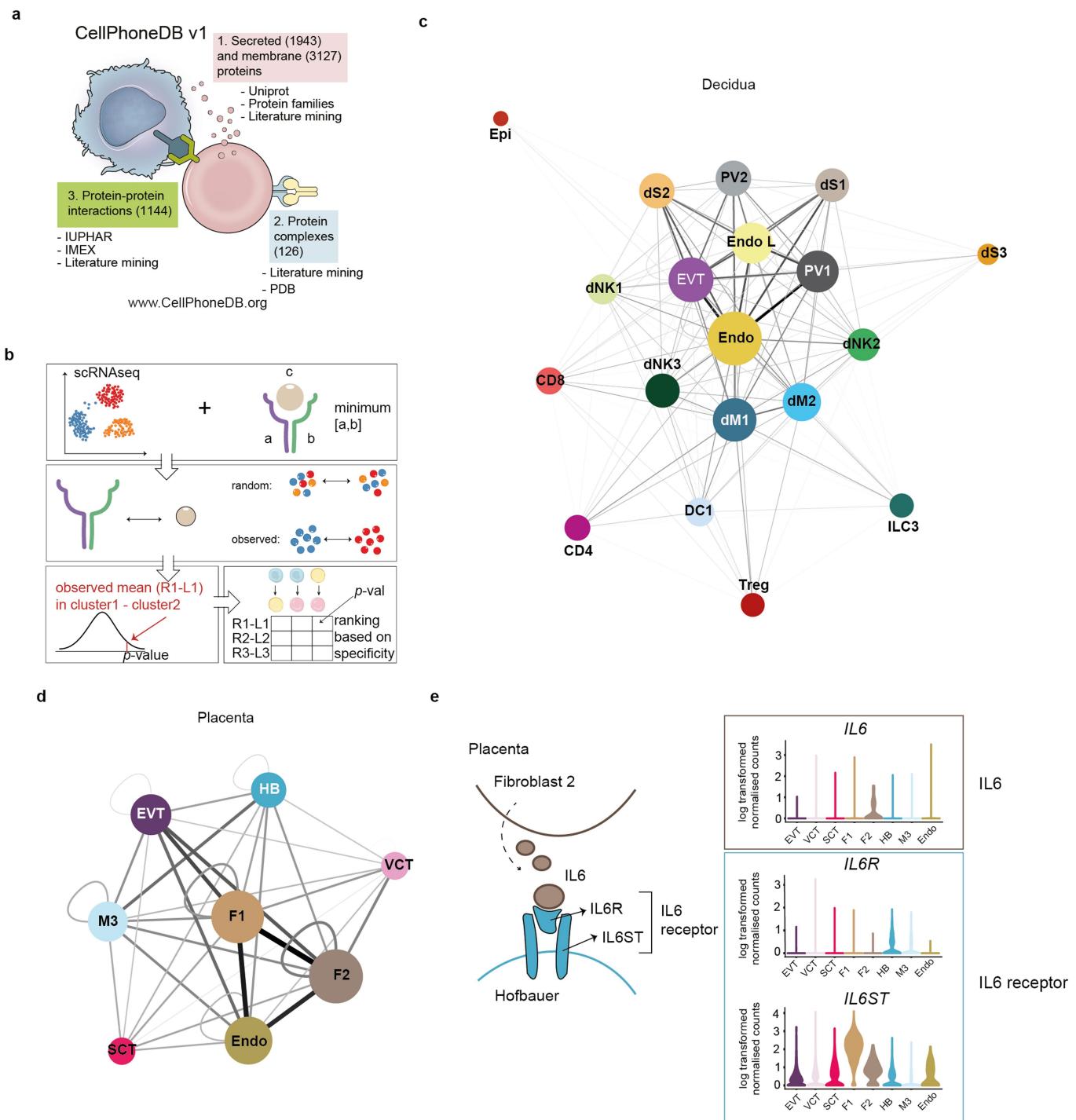


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Overview of droplet and Smart-seq2 datasets.

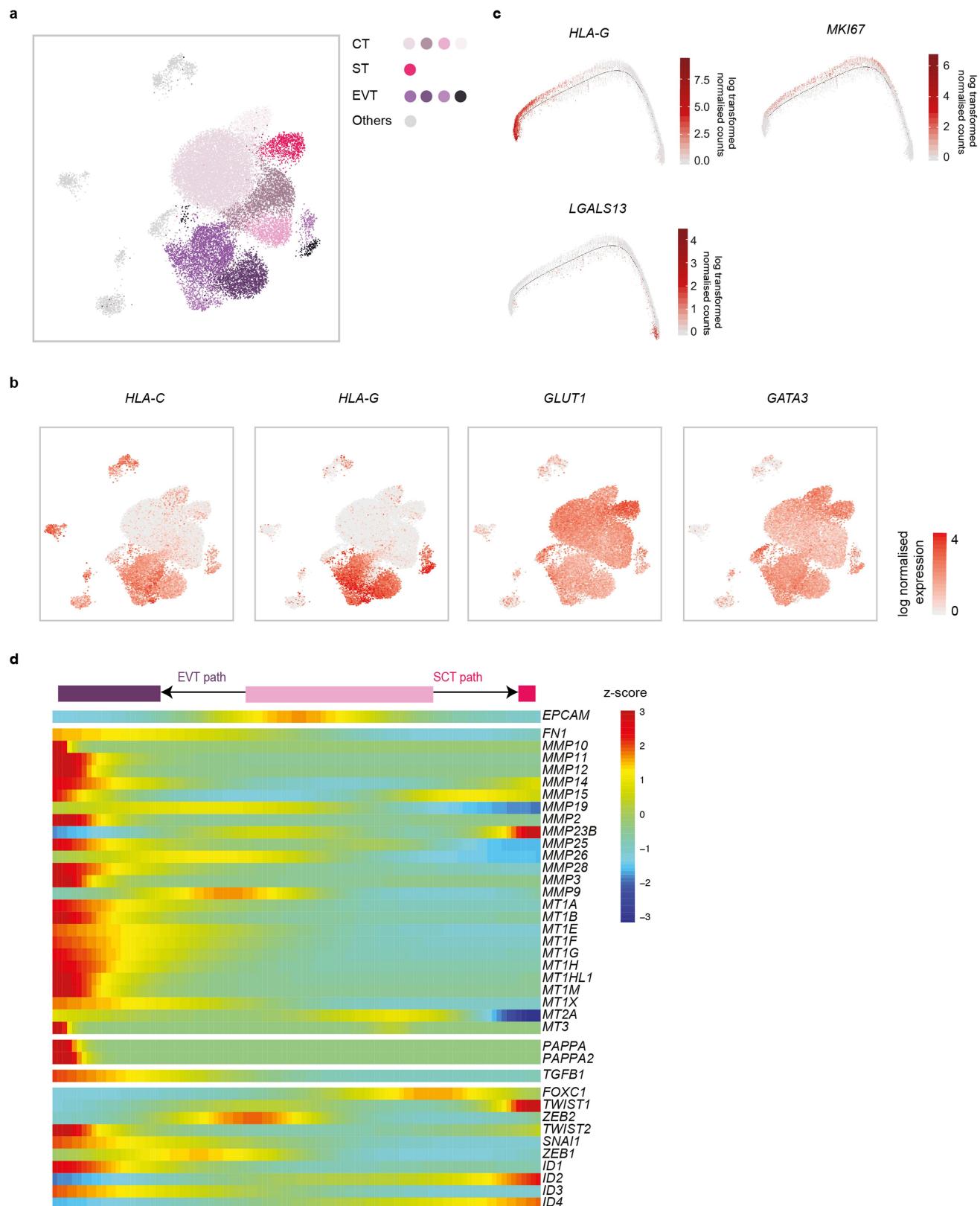
a, UMAP plot showing the integration of the Smart-seq2 and droplet-based dataset and the log-transformed expression of *MKI67* (which marks proliferating cells). **b**, UMAP plots showing the separate and more-detailed integration analysis of the cells from cluster 14 (perivascular cells), cluster 19 (endothelial cells) and cluster 25 (epithelial cells). Clusters are labelled as in Fig. 1c. **c**, UMAP visualization of T cell clusters obtained by integrating Smart-seq2 and droplet-based T cells subpopulations (clusters 4, 8, 10 and 15) from Fig. 1c. Cells are coloured by the tissue of origin (top) and the identified clusters (bottom). **d**, Heat map showing the z-score of the mean log-transformed, normalized counts for each cluster of selected marker genes used to annotate clusters. For a more extensive

set of genes, see Supplementary Table 2. Adjusted *P* value < 0.1; Wilcoxon rank-sum test with Bonferroni correction. NK, natural killer cells; NKp, proliferating natural killer cells; MO, monocytes; Granulo, granulocytes; T_{reg}, regulatory T cells; GD, $\gamma\delta$ T cells; CD8c, cytotoxic CD8⁺ T cells; Plasma, plasma cells. **e**, log-likelihood differences between assignment to fetal versus assignment to maternal origin of cells, on the basis of single nucleotide polymorphism calling from the droplet RNA-seq data. Cells are coloured by their assignment as determined by demuxlet. For this figure, we used *n* = 5 placentas, *n* = 6 decidua and *n* = 4 blood individuals. **f**, UMAP visualization of the log-transformed, normalized expression of selected marker genes of the M3 subpopulation.



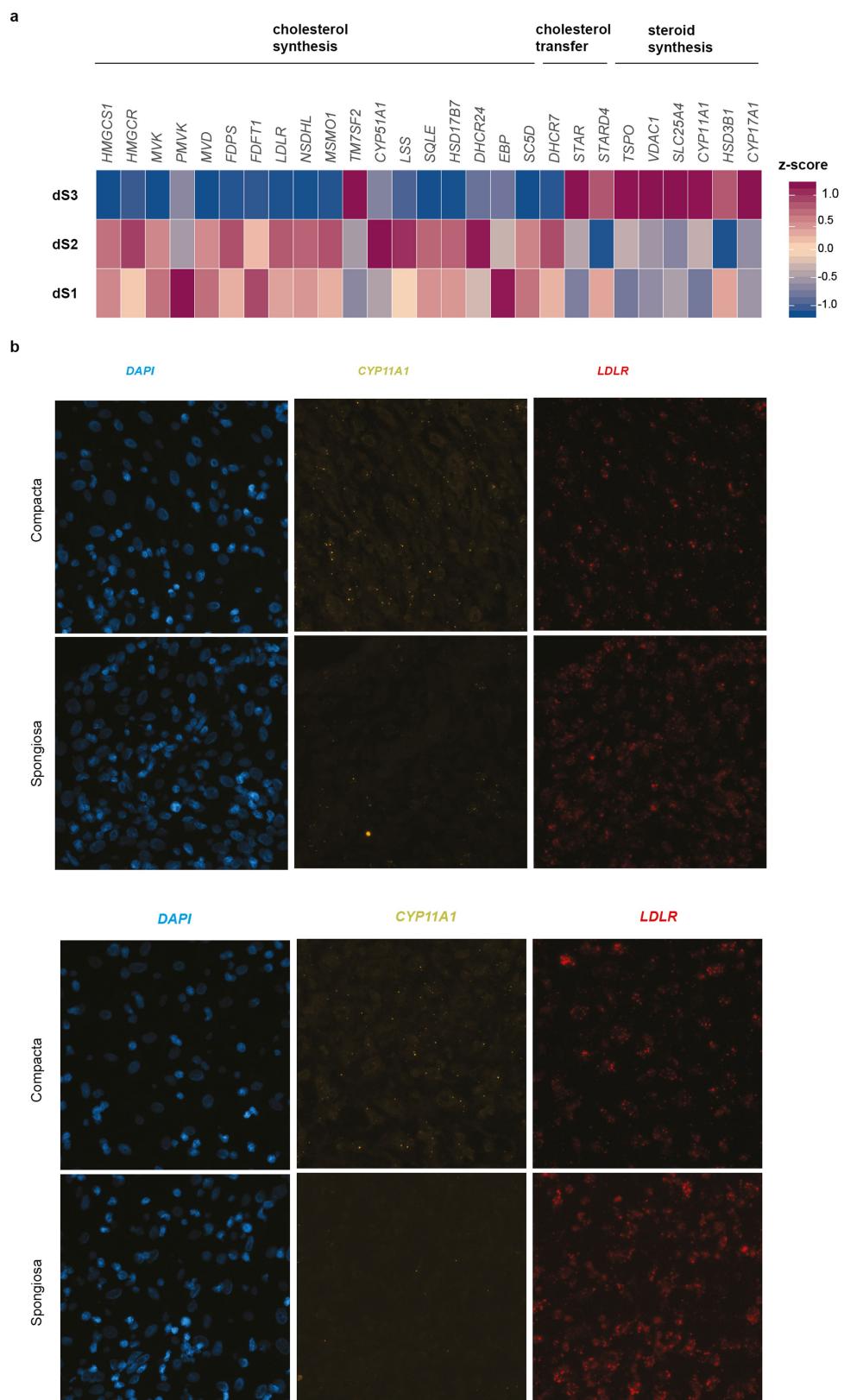
Extended Data Fig. 4 | Cell-cell communication networks in the maternal-fetal interface using CellPhoneDB. **a**, Information aggregated within www.CellPhoneDB.org. **b**, Statistical framework used to infer ligand-receptor complex specific to two cell types from single-cell transcriptomics data. Predicted *P* values for a ligand-receptor complex across two cell clusters are calculated using permutations, in which cells are randomly re-assigned to clusters (see Methods). **c**, Networks visualizing potential specific interactions in the decidua, in which nodes are clusters (cell types) and edges represent the number of significant ligand-receptor pairs. The network was created for edges with more than 30 interactions and the network layout was set to force-directed layout. Only droplet

data were considered for the CellPhoneDB analysis ($n=6$ decidua). **d**, Networks visualizing potential specific interactions in the placenta, in which nodes are clusters and edges represent the number of significant ligand-receptor pairs. The network layout was set to force-directed layout. Only droplet data were considered for the analysis ($n=5$ placentas). **e**, An example of significant interactions identified by CellPhoneDB. Violin plots show log-transformed, normalized expression levels of the components of the IL6-IL6R complex in placental cells. IL6 expression is enriched in the fibroblast 2 cluster (F2; dark brown in **d**) and the two subunits of the IL6 receptors (IL6R and IL6ST) are co-expressed in Hofbauer cells.



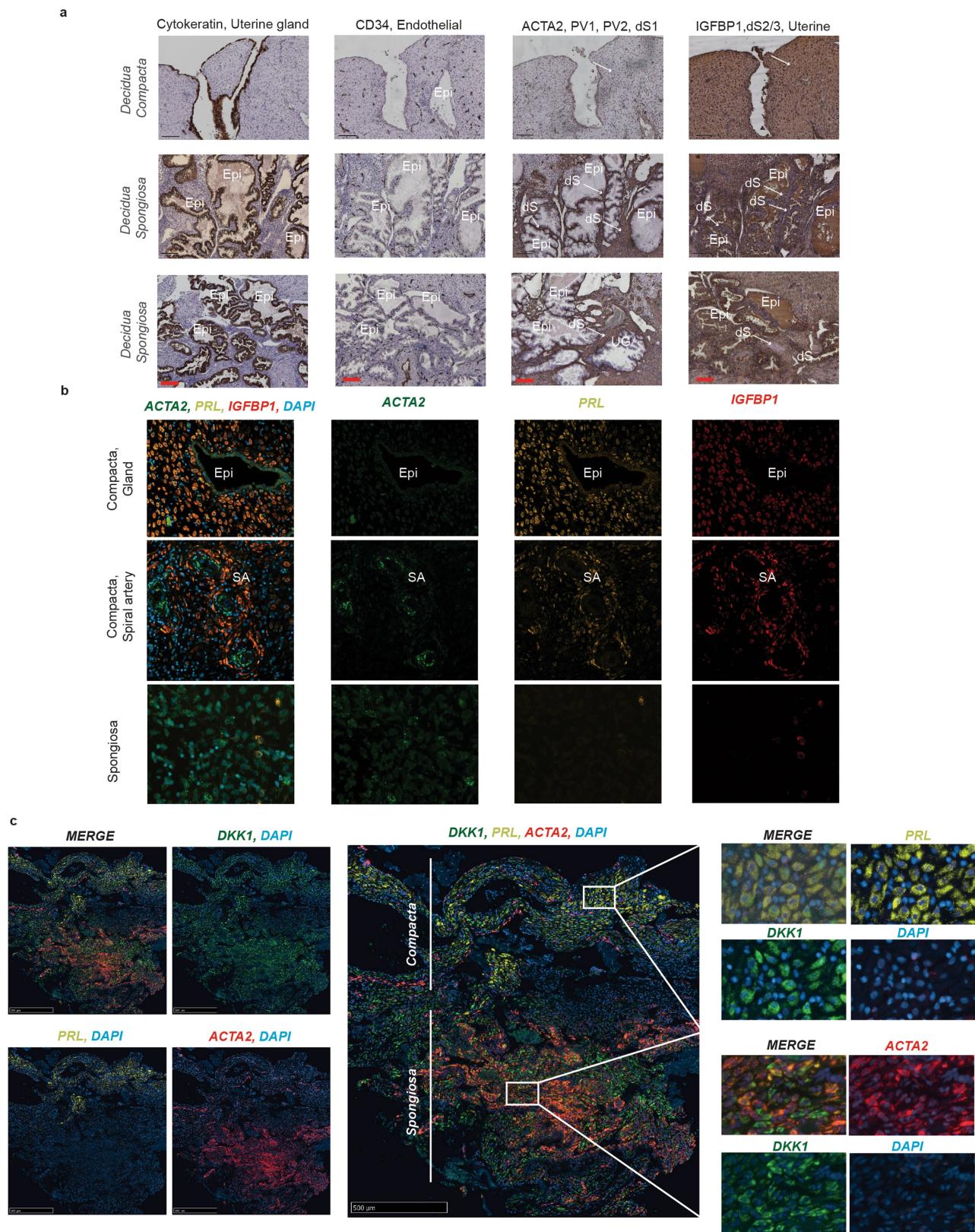
Extended Data Fig. 5 | Trophoblast analysis. **a**, UMAP visualization of the integrated analysis of the trophoblast subpopulations that were used for pseudotime analysis, including the enriched EPCAM⁺ and HLA-G⁺ cells (see Methods). Cells that were excluded from the pseudotime analysis are coloured in grey ($n=5$ placentas, $n=11$ decidua). **b**, UMAP visualization of the log-transformed, normalized expression of selected

canonical trophoblast marker genes ($n=5$ placentas). **c**, Visualization of log-transformed, normalized expression of *HLA-G*, *MKI67* and *LGALS13* across trophoblast differentiation. **d**, Heat map showing genes that are involved in the epithelial–mesenchymal transition, identified as varying significantly as EVT differentiate (q value < 0.1 , likelihood ratio test, P values were adjusted for the false discovery rate).



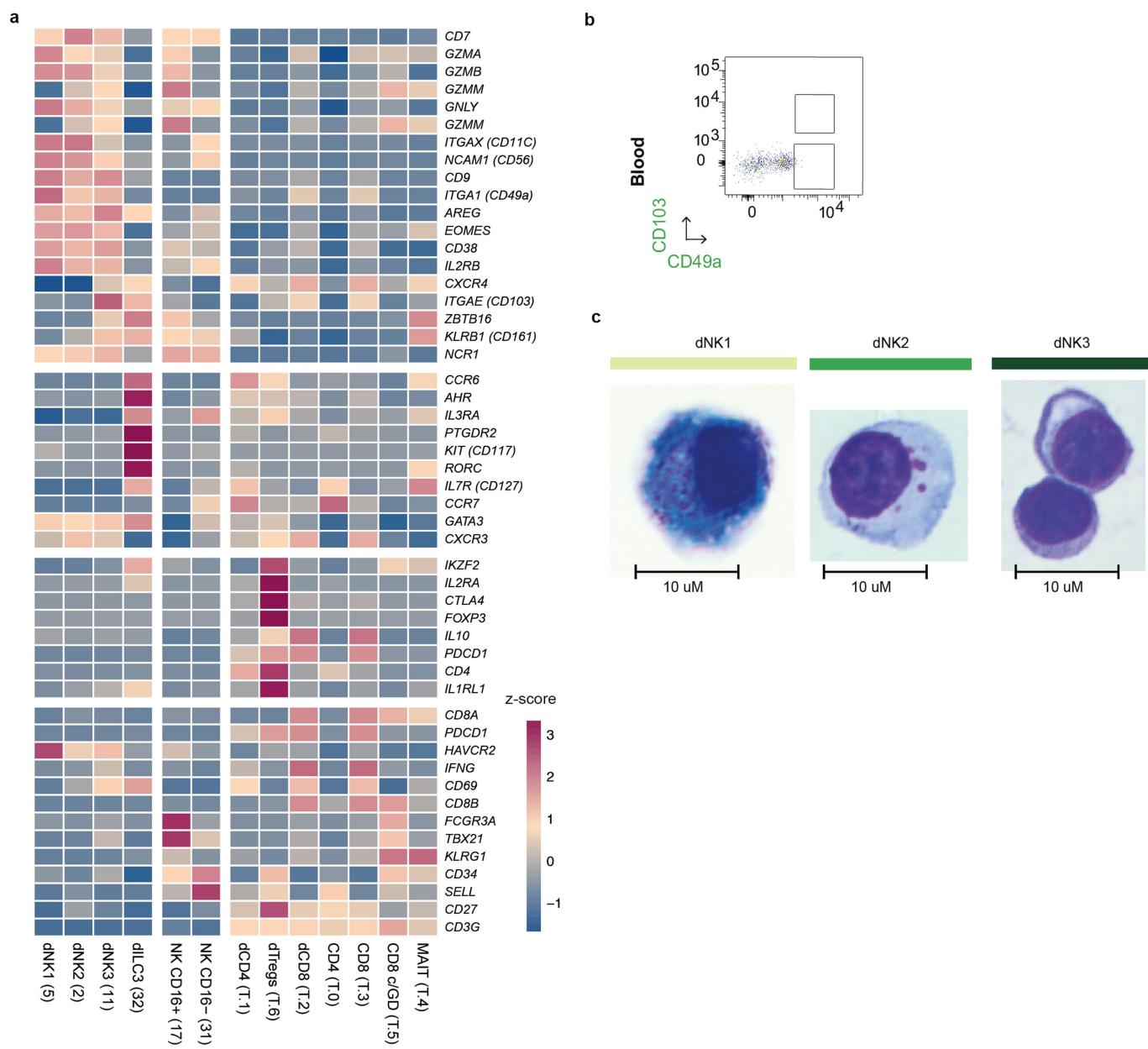
Extended Data Fig. 6 | Steroid synthesis. **a**, Heat map showing relative expression of enzymes involved in cholesterol and steroid synthesis in the three stromal subsets ($n = 11$ decidua). **b**, Multiplexed smFISH in two decidua parietalis sections from two different individuals, showing

an enrichment of CYP11A1 expression in the decidua compacta. Section stained by CYP11A1, LDLR and DAPI. Images are shown at $40\times$ magnification. A high resolution is needed to detect differences between the sections ($n = 2$ individuals).



Extended Data Fig. 7 | In situ staining for the different stromal cells.
a, Immunohistochemistry of decidual serial sections stained for cytokeratin (uterine glands), CD34 (endothelial cells), ACTA2 (perivascular populations and dS1) and IGFBP1 (stromal cells and glandular secretions) ($n = 2$ biological replicates). ACTA2⁺ stromal cells are confined to the stromal cells of the deeper decidua spongiosa, whereas stromal cells in the decidua compacta are ACTA2⁻. IGFBP1⁺ stromal cells are enriched in the decidua compacta, whereas stromal cells around the glands in the decidua spongiosa are IGFBP1⁻. Glandular secretions are

IGFBP1⁺. **b**, Multiplexed smFISH for a decidua parietalis section showing the two decidual layers. ACTA2, dS1 population confined to decidua spongiosa; IGBP1 and PRL, dS2 and dS3 populations confined to decidua compacta. Samples shown are from a different individual than samples shown in Fig. 4d ($n = 2$ biological replicates). **c**, Multiplexed smFISH for a decidua parietalis section showing the two decidual layers. DKK1, decidual stromal marker; ACTA2, dS1 population confined to decidua spongiosa; PRL, dS3 population confined to decidua compacta ($n = 1$ biological replicate).

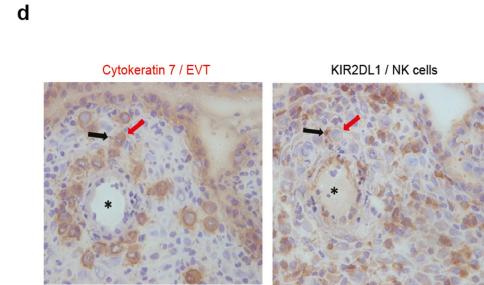
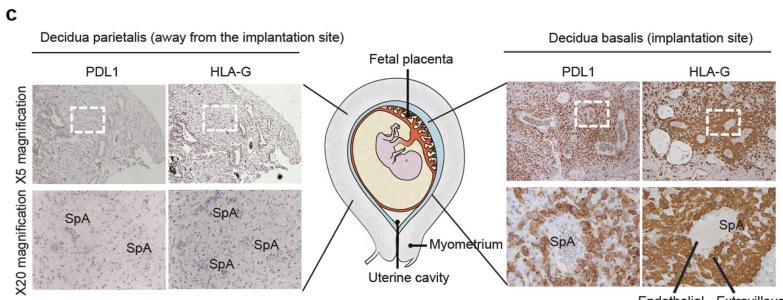
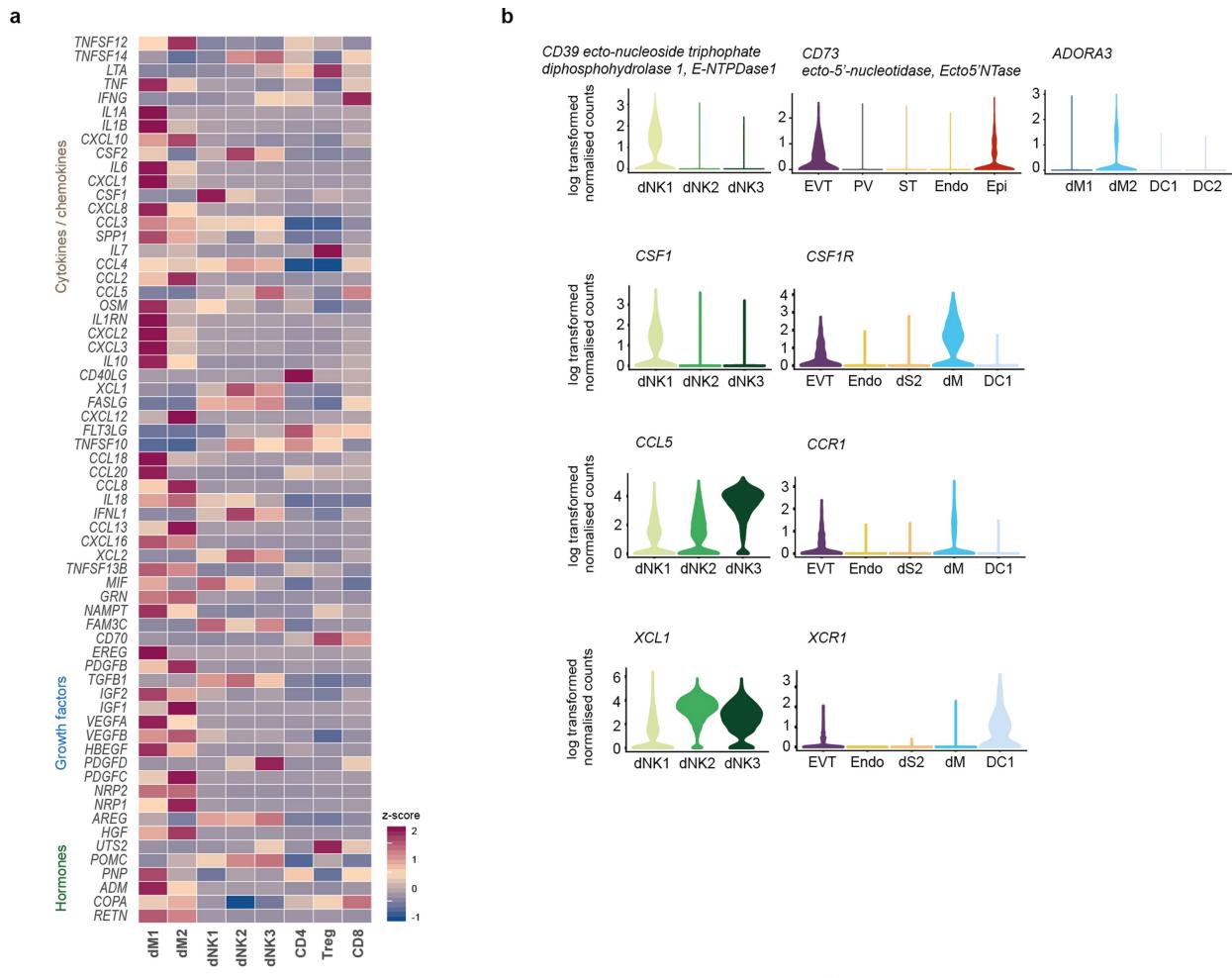


Extended Data Fig. 8 | Lymphocyte populations in the decidua.

a, Heat map showing z-scores of the mean log-transformed, normalized expression of selected genes in the lymphocyte populations. Proliferating dNK cells (dNKp) are excluded from the analysis ($n=11$ decidua).

b, FACS gating strategy in Fig. 5 applied in matched blood. Matched blood

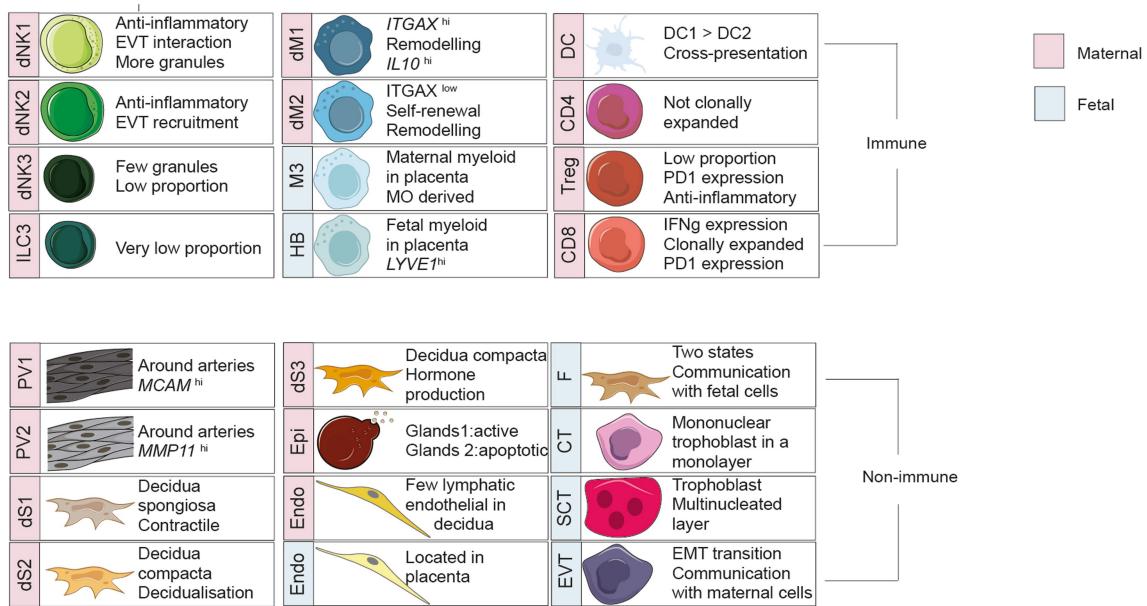
for the sample shown in Fig. 5 ($n=2$ biological replicates). **c**, Morphology of dNK1, dNK2 and dNK3 subsets by Giemsa-Wright stain after cytopsin (representative data from 1 of $n=2$ biological replicates are shown). Scale bar, $10 \mu\text{m}$.



Extended Data Fig. 9 | Expression of ligands and receptors at the maternal-fetal interface. **a**, Heat map showing z-scores of the mean log-transformed, normalized expression of genes annotated as cytokines, growth factors, hormones and angiogenic factors with a log-mean > 0.1 in the selected decidual immune populations ($n = 11$ decidua). **b**, Violin plots showing log-transformed, normalized expression levels of selected ligands expressed in the three dNK cells and their corresponding receptors expressed on other decidual cells and EVT (*CD39*, *CD73*, *ADORA3*, *CSF1*, *CSF1R*, *CCL5*, *CCR1*, *XCL1* and *XCR1*; $n = 11$ decidua, $n = 5$ placentas). **c**, Immunohistochemistry images of serial decidual sections stained for the EVT marker HLA-G and the inhibitory ligand PDL1. Bottom panels

shown the areas in white boxes in the top panels at higher power. HLA-G⁺ cells are only present at the site of placentation (decidua basalis) and are absent elsewhere (decidua parietalis). SpA, spiral arteries. The EVT is strongly PDL1⁺. We show representative data from one individual of $n = 5$ biological replicates. **d**, Immunohistochemistry images of decidual serial sections of the decidual implantation site (at 10 weeks of gestation), stained for the trophoblast cell marker, cytokeratin-7 (red arrow) and the inhibitory receptor KIR2DL1 on a natural killer cell (black arrow). The asterisk marks the lumen of a spiral artery that supplies the conceptus. We show representative data from one individual of $n = 5$ samples).

a



Extended Data Fig. 10 | Encyclopaedia of cells at the maternal–fetal interface. a, Summary of populations from our scRNA-seq data. Blue, fetal; red, maternal.

De novo NAD⁺ synthesis enhances mitochondrial function and improves health

Elena Katsyuba¹, Adrienne Mottis¹, Marika Zietak^{2,3}, Francesca De Franco⁴, Vera van der Velpen⁵, Karim Gariani^{1,11}, Dongryeol Ryu^{1,12}, Lucia Cialabrini⁶, Olli Matilainen^{1,13}, Paride Liscio⁴, Nicola Giacchè⁴, Nadine Stokar-Regenscheit^{7,14}, David Legouis^{8,9}, Sophie de Seigneux^{9,10}, Julijana Ivanisevic⁵, Nadia Raffaelli⁶, Kristina Schoonjans², Roberto Pellicciari^{4*} & Johan Auwerx^{1*}

Nicotinamide adenine dinucleotide (NAD⁺) is a co-substrate for several enzymes, including the sirtuin family of NAD⁺-dependent protein deacetylases. Beneficial effects of increased NAD⁺ levels and sirtuin activation on mitochondrial homeostasis, organismal metabolism and lifespan have been established across species. Here we show that α -amino- β -carboxymuconate- ϵ -semialdehyde decarboxylase (ACMSD), the enzyme that limits spontaneous cyclization of α -amino- β -carboxymuconate- ϵ -semialdehyde in the de novo NAD⁺ synthesis pathway, controls cellular NAD⁺ levels via an evolutionarily conserved mechanism in *Caenorhabditis elegans* and mouse. Genetic and pharmacological inhibition of ACMSD boosts de novo NAD⁺ synthesis and sirtuin 1 activity, ultimately enhancing mitochondrial function. We also characterize two potent and selective inhibitors of ACMSD. Because expression of ACMSD is largely restricted to kidney and liver, these inhibitors may have therapeutic potential for protection of these tissues from injury. In summary, we identify ACMSD as a key modulator of cellular NAD⁺ levels, sirtuin activity and mitochondrial homeostasis in kidney and liver.

Increasing NAD⁺ levels activate the sirtuins and have a positive effect on metabolism in different model organisms^{1–4}. Given the beneficial effects of replenished NAD⁺ pools, there is an intense search for strategies to increase intracellular NAD⁺ by limiting NAD⁺ consumption or increasing NAD⁺ production⁵. NAD⁺ can be produced via salvage pathways or by de novo synthesis. De novo NAD⁺ synthesis starts from the amino acid tryptophan⁶. The formation of unstable α -amino- β -carboxymuconate- ϵ -semialdehyde (ACMS) constitutes a branching point of this pathway (Extended Data Fig. 1a). ACMS undergoes either cyclization, forming the NAD⁺ precursor quinolinic acid, or total oxidation to CO₂ and H₂O. Whereas the cyclization of ACMS occurs spontaneously, the transformation of ACMS to α -amino- β -muconate- ϵ -semialdehyde (AMS) is catalysed by ACMSD, which determines the proportion of ACMS able to undergo cyclization and produce NAD⁺. ACMSD is conserved across species, with mouse, rat and *C. elegans* (ACSD-1, Y71D11A.3) orthologues showing 85, 85 and 48% similarity, respectively, to the human protein⁷. On the basis of this sequence conservation, we initially characterized the function of ACSD-1 in *C. elegans*.

acsd-1 controls NAD⁺ levels in *C. elegans*

The newly generated *C. elegans* *acsd-1*::GFP reporter strain showed expression of *acsd-1* in the majority of tissues throughout development

and adulthood (Extended Data Fig. 1b, c). *acsd-1* RNA-mediated interference (RNAi)—by feeding worms with HT115 *Escherichia coli* that expresses *acsd-1* RNA—decreased *acsd-1* transcript levels by 46% in wild-type (N2) worms, and by 78% in *rrf-3*(pk1426) mutant worms, which are hypersensitive to RNAi (Extended Data Fig. 1d). *acsd-1* RNAi resulted in complete loss of function (LOF) of ACSD-1 enzymatic activity in *rrf-3* mutants and 70% loss of ACSD-1 activity in N2 worms (Extended Data Fig. 1e).

It has long been postulated^{8,9}, on the basis that they do not possess a quinolinate phosphoribosyltransferase (QPRT) orthologue with obvious sequence similarities (Extended Data Fig. 1a), that *C. elegans* cannot synthesize NAD⁺ de novo and therefore rely on pre-formed pyridine rings to produce NAD⁺. This was recently disproved, as uridine monophosphate synthetase replaces this function of QPRT in the nematode¹⁰. We detected QPRT-like enzymatic activity in both N2 and *rrf-3* worms (Extended Data Fig. 1f), and showed that tryptophan dose-dependently increased NAD⁺ levels (Fig. 1a). Finally, *acsd-1* RNAi increased NAD⁺ content 1.2-fold (Fig. 1b).

Increases in NAD⁺ are known to extend lifespan of worms^{11,12}. *acsd-1* RNAi did not affect N2 lifespan in basal conditions (Extended Data Fig. 1g), but survival of *rrf-3* mutants was significantly increased (Extended Data Fig. 1h). The lifespan-enhancing effect therefore seems to depend on the extent of *acsd-1* downregulation. Consistent with

¹Laboratory of Integrative and Systems Physiology, Interfaculty Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. ²Laboratory of Metabolic Signaling, Interfaculty Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. ³Institute of Animal Reproduction and Food Research, Polish Academy of Sciences, Olsztyn, Poland. ⁴TES Pharma, Loc. Taverne, Corciano, Italy. ⁵Metabolomics Platform, Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland. ⁶Department of Agricultural, Food and Environmental Sciences, Polytechnic University of Marche, Ancona, Italy. ⁷Histology Core Facility, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. ⁸Intensive Care Unit, Department of Anaesthesiology, Pharmacology and Intensive Care, University Hospital of Geneva, Geneva, Switzerland. ⁹Laboratory of Nephrology, Department of Internal Medicine Specialties and Department of Cell Physiology and Metabolism, University of Geneva, Geneva, Switzerland. ¹⁰Service of Nephrology, Department of Internal Medicine Specialties, University Hospital of Geneva, Geneva, Switzerland. ¹¹Present address: Service of Endocrinology, Diabetes, Hypertension and Nutrition, Geneva University Hospitals, Geneva, Switzerland. ¹²Present address: Molecular and Integrative Biology Lab, Healthy Aging-Korean Medical Research Center, Department of Korean Medical Science, School of Korean Medicine, Pusan National University, Yangsan, South Korea. ¹³Present address: Institute of Biotechnology, University of Helsinki, Helsinki, Finland. ¹⁴Present address: Roche Pharma Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, F. Hoffmann-La Roche, Basel, Switzerland. *e-mail: rpellicciari@tespharma.com; admin.auwerx@epfl.ch

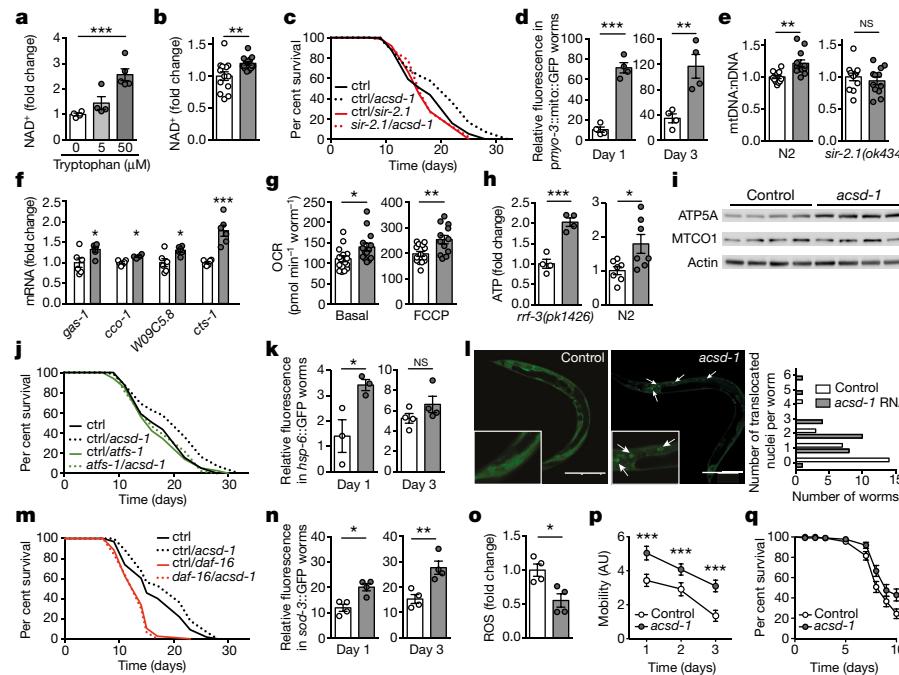


Fig. 1 | acsd-1 LOF increases NAD⁺ levels, improves mitochondrial function and increases lifespan through de novo synthesis in *C. elegans*. **a, b**, Increase in NAD⁺ levels in worms with tryptophan supplementation (**a**; $n = 4$ (0 and 5 μ M), $n = 5$ (50 μ M)) and feeding with control (empty vector) or *acsd-1* RNAi (**b**; $n = 14$). Each n represents a pool of $\sim 1,000$ worms. **c**, Epistasis of *acsd-1* with *sir-2.1* RNAi. Control versus ctrl + *acsd-1* RNAi, $P < 0.0001$; ctrl + *sir-2.1* RNAi versus *sir-2.1* RNAi + *acsd-1* RNAi, not significant. **d**, GFP signal in the reporter strain, expressing a mitochondria-targeted GFP in muscle at day 1 and 3 of adulthood ($n = 4$ pools of 20 worms). **e**, mtDNA:nDNA ratio in wild-type (N2) and *sir-2.1(ok434)* mutant worms ($n = 12$ worms) with control or *acsd-1* RNAi. **f, g**, Changes in expression of mRNA encoding mitochondrial proteins (**f**; $n = 6$ pools of ~ 600 worms) and oxygen consumption rate (OCR) in basal and uncoupled conditions (**g**; $n = 14$ pools of 10 worms) in worms fed with control or *acsd-1* RNAi. **h**, ATP content in *rff-3(pk1426)* and N2 worms fed with control (empty vector) or *acsd-1* RNAi ($n = 4$ or 7 pools of ~ 100 worms, respectively). **i**, Expression of OXPHOS subunits encoded by nDNA (ATP5A) and mtDNA (MTCO1) with control or *acsd-1* RNAi. Each lane represents a pool of ~ 600 worms. **j**, Epistasis of *acsd-1* with the UPR^{mt} regulator *atfs-1* RNAi. Control RNAi versus ctrl + *acsd-1* RNAi, $P < 0.001$; ctrl + *atfs-1* RNAi versus *atfs-1* + *acsd-1* RNAi, not significant. **k**, GFP

reported data¹³, and with the effects of tryptophan supplementation on NAD⁺ concentration, tryptophan also extended worm lifespan, and *acsd-1* RNAi did not result in a significant further extension beyond the effects of tryptophan alone (Extended Data Fig. 1i). Although the effects of *sir-2.1* overexpression on lifespan are disputed^{14,15}, *sir-2.1* has been shown to have a role in NAD⁺-mediated lifespan extension^{11,12}. Accordingly, we found that *sir-2.1* was required for the longevity effect of *acsd-1* RNAi (Fig. 1c), supporting the role of *sir-2.1* in NAD⁺-dependent lifespan regulation. As ACSD-1 is involved in several molecular pathways, the effects upon loss of its function may not be mediated exclusively by changes in NAD⁺ levels, but could also arise partly from changes in the levels of products of ACSD-1 enzymatic activity, such as AMS or picolinic acid. The contribution of these downstream factors remains to be determined.

Sirtuin activation is associated with enhanced mitochondrial function^{1,2}. *acsd-1* RNAi in the *myo-3::GFP(mit)* and *ges-1::GFP(mit)* reporter strains—which express mitochondria-targeted GFP in muscle and intestinal cells, respectively—robustly increased mitochondrial content in both tissues (Fig. 1d, Extended Data Fig. 1j). Measurement of the ratio of mitochondrial DNA (mtDNA) to nuclear DNA (nDNA) confirmed the increase in mitochondria upon *acsd-1* RNAi in wild-type worms, but not in *sir-2.1(ok434)* mutants (Fig. 1e). Furthermore, *acsd-1*

signal in *hsp-6::GFP* reporter strain fed with control or *acsd-1* RNAi at day 1 and 3 of adulthood (day 1, $n = 3$; day 3, $n = 4$; each n represents a pool of 20 worms). **l**, Nuclear translocation of DAF-16. Arrowheads indicate DAF-16 accumulation within nuclei. The graph represents the distribution of worms treated with control or *acsd-1* RNAi in which DAF-16 has translocated to the nuclei ($n = 25$ worms). Scale bar, 100 μ m. **m**, Epistasis of *acsd-1* with *daf-16* RNAi. Control RNAi versus ctrl + *acsd-1* RNAi, $P < 0.001$; ctrl + *daf-16* RNAi versus *daf-16* + *acsd-1* RNAi, not significant. **n**, GFP signal in *sod-3::GFP* reporter worms at day 1 and 3 of adulthood treated with control and *acsd-1* RNAi ($n = 4$ pools of 20 worms). **o**, ROS in worms exposed to control or *acsd-1* RNAi ($n = 4$ pools of 20 worms). **p, q**, Mobility (**p**) and survival (**q**; $P = 0.003$) in N2 worms exposed to 4 mM paraquat starting from L4 stage, treated with control or *acsd-1* RNAi throughout their life ($n = 100$ worms). All worm assays were performed at 20 °C and repeated at least once. Data are mean \pm s.e.m. * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$; NS, not significant. P values calculated using one-way ANOVA (**a**), two-tailed *t*-test (**b, d–h, k, n–p**) or log-rank test (**c, j, m, q**). For gel source images, see Supplementary Fig. 1. For individual P values, see Source Data. For lifespan values, see Extended Data Table 1. AU, arbitrary units.

LOF increased transcript levels of many mitochondrial genes, basal and maximal respiration, mitochondrial complex II levels and ATP content (Fig. 1f–h, Extended Data Fig. 1k). The mitochondrial network was also more extensive and interconnected in worms treated with *acsd-1* RNAi (Extended Data Fig. 1l).

acsd-1 LOF activates mitochondrial stress defence

Notably, *acsd-1* RNAi induced expression of proteins associated with oxidative phosphorylation (OPXPHOS) and proteins encoded by nDNA, such as H28O16.1 (an orthologue of mammalian ATP5A), whereas expression of OXPHOS components encoded by mtDNA—such as MTCE.26 (an orthologue of mammalian MTCO1)—was unchanged (Fig. 1i). Changes in the ratio of proteins encoded by mtDNA versus nDNA are a hallmark of mitonuclear protein imbalance, which can be induced by increasing NAD⁺ concentrations¹¹ and which is associated with the activation of the mitochondrial unfolded protein response (UPR^{mt})¹⁶. LOF of either *atfs-1* or *ubl-5*—which are essential UPR^{mt} genes^{17,18}—attenuated the increased longevity resulting from *acsd-1* RNAi alone (Fig. 1j, Extended Data Fig. 1m), indicating that the UPR^{mt} is required for the lifespan extension effect. By using *hsp-6::GFP*, a reporter for a mitochondrial chaperone orthologous to mammalian mtHsp70, we observed a robust activation of the UPR^{mt}

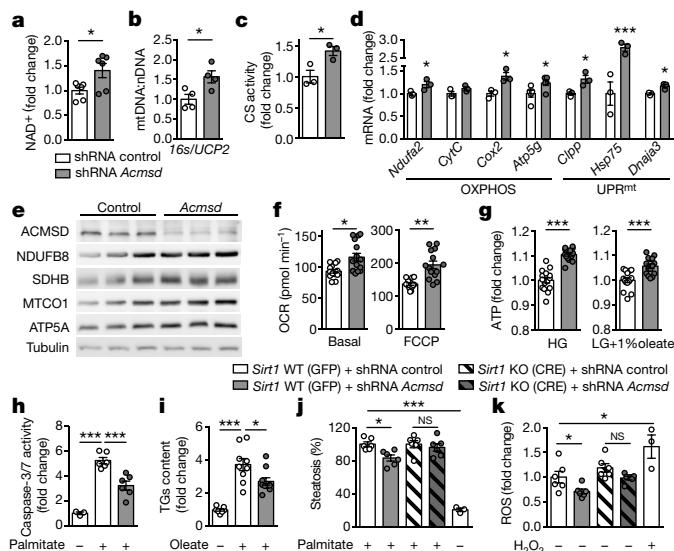


Fig. 2 | Pathways activated by *Acmsd* knockdown in worms are conserved in mammalian cells. **a–f**, Mouse primary hepatocytes obtained from C57BL/6J mice were transduced with an adenovirus encoding either control shRNA or shRNA directed against *Acmsd* for 48 h. Change in NAD⁺ levels ($n=6$) (**a**), mtDNA:nDNA ratio ($n=4$) (**b**), citrate synthase (CS) activity ($n=3$) (**c**), expression of OXPHOS and UPR^{mt} genes ($n=3$) (**d**), protein expression of respiratory complex subunits (**e**) and oxygen consumption in basal and uncoupled (2 μ M *p*-trifluoromethoxy carbonyl cyanide phenyl hydrazone (FCCP)) conditions ($n=15$) (**f**). **g**, Change in ATP levels in mouse primary hepatocytes grown in both medium with high glucose (HG) or low glucose (LG) supplemented with 1% oleate treated with control shRNA versus *Acmsd* shRNA ($n=15$). **h**, Apoptosis evaluated by caspase-3/7 activity after 24-h exposure of mouse primary hepatocytes to 0.75 mM palmitate (no palmitate, $n=3$; palmitate, $n=6$). **i**, Triglyceride (TGs) content after 24-h exposure to 0.5 mM oleate in AML12 cells transduced with an adenovirus encoding either control or *Acmsd* shRNA ($n=9$). **j, k**, Primary hepatocytes extracted from a *Sirt1*^{L2/L2} mouse transduced with an adenovirus encoding either GFP (wild-type) or Cre recombinase (to generate *Sirt1* knockout (KO) primary hepatocytes). **j**, Steatosis in hepatocytes treated with control or *Acmsd* shRNA after 24-h exposure to 0.75 mM palmitate (no palmitate, $n=3$; palmitate, $n=6$). **k**, ROS content in hepatocytes exposed to control shRNA versus *Acmsd* shRNA. The positive control is treated with 550 μ M H₂O₂ (no H₂O₂, $n=3$; H₂O₂, $n=6$). Data are mean \pm s.e.m.; n represents number of biologically independent samples. All experiments performed independently at least twice. * $P\leq 0.05$, ** $P\leq 0.01$, *** $P\leq 0.001$. P values calculated using two-tailed *t*-test (**a–d, f–i**) or two-way ANOVA (**j–k**). For gel source images, see Supplementary Fig. 1. For individual P values, see Source Data.

upon *acsd-1* RNAi (Fig. 1k). This stress response was specific, as there was no activation of the unfolded protein response in the endoplasmic reticulum or of the cytosolic heat shock response in *hsp-4::GFP* and *hsp-16.2::GFP* reporter strains, respectively, after *acsd-1* LOF (Extended Data Fig. 1n–o). Several UPR^{mt}-related transcripts were induced upon *acsd-1* RNAi (Extended Data Fig. 1p).

In *C. elegans*, oxidative stress defence is launched when DAF-16—an orthologue of forkhead box O transcription factor (FOXO)—is translocated into the nucleus¹⁹, where it induces the mitochondrial superoxide dismutase²⁰ *sod-3*. Supplementing *C. elegans* with NAD⁺ was reported to activate oxidative stress defence and to improve resistance to reactive oxygen species (ROS) in a *sir-2.1*- and *daf-16*-dependent fashion¹². Consistent with this mechanism, DAF-16 translocated into nuclei upon *acsd-1* LOF (Fig. 1l). The increased lifespan as a result of *acsd-1* RNAi was lost upon *daf-16* LOF (Fig. 1m). Consistent with increased nuclear localization of DAF-16, *acsd-1* LOF increased the GFP signal in *sod-3::GFP* reporter worms as well as *sod-3* expression (Fig. 1n, Extended Data Fig. 1q). As a consequence, worms with *acsd-1* knockdown had reduced levels of ROS, were more active and lived longer when exposed to paraquat, a known ROS inducer (Fig. 1o–q). The increased survival under paraquat was

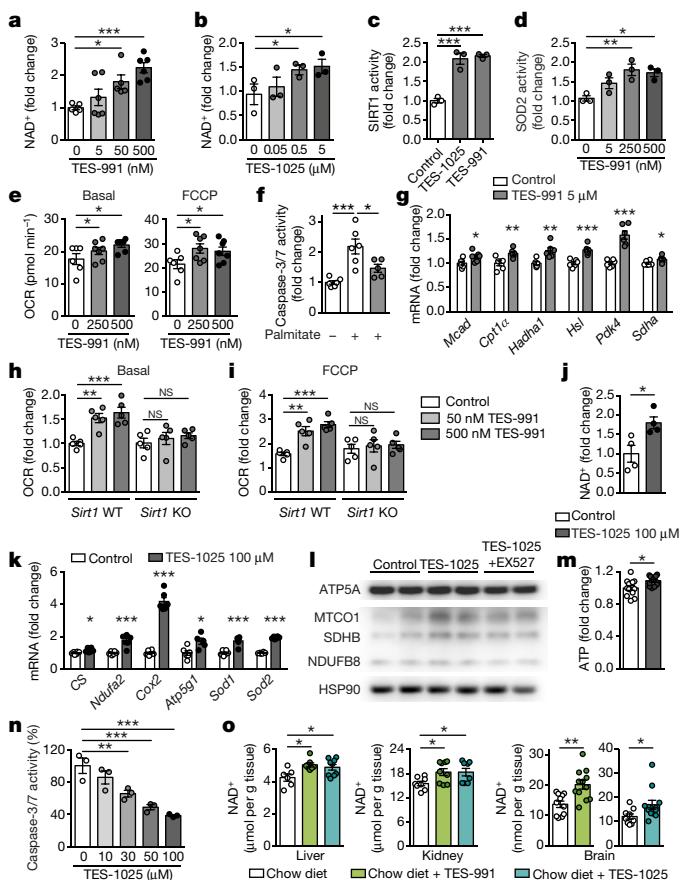


Fig. 3 | Pharmacological inhibition of ACMSD has similar effects to genetic downregulation. **a, b**, NAD⁺ levels in mouse primary hepatocytes treated for 24 h with vehicle (dimethyl sulfoxide, DMSO), TES-991 ($n=6$) (**a**) or TES-1025 ($n=3$) (**b**). **c**, SIRT1 activity in primary hepatocytes treated for 24 h with vehicle or 500 nM TES-1025 or TES-991 ($n=3$). **d–f**, SOD2 activity ($n=3$) (**d**), oxygen consumption in basal and uncoupled conditions (ctrl, $n=6$; TES-991, $n=7$) (**e**) and apoptosis rate after 36 h exposure to 0.75 mM palmitate (ctrl, $n=6$; TES-991, $n=5$) (**f**) in primary hepatocytes treated for 24 h with vehicle or TES-991. **g**, mRNA levels of fatty acid oxidation (FAO) genes in primary hepatocytes treated with vehicle or 500 nM TES-991 after 6 h exposure to 0.33 mM palmitate and 0.66 mM oleate ($n=6$). **h, i**, Primary hepatocytes from a *Sirt1*^{L2/L2} mouse transduced with an adenovirus encoding GFP (*Sirt1* WT) or Cre recombinase (*Sirt1* KO). FAO after 24-h treatment with vehicle or TES-991 under basal (**h**) or uncoupled (**i**) conditions ($n=5$). **j–m**, Changes in NAD⁺ ($n=4$) (**j**), mRNA levels of mitochondrial and oxidative stress defence genes ($n=6$) (**k**), protein levels of OXPHOS subunits (**l**) and ATP content (ctrl, $n=12$; TES-1025, $n=13$) (**m**) in HK-2 cells upon treatment with 100 μ M TES-1025, or TES-1025 in combination with 10 μ M SIRT1 inhibitor EX527, for 24 h. **n**, Apoptosis in HK-2 cells 16 h after addition of 50 μ M cisplatin. TES-1025 was added 1 h before cisplatin ($n=3$). **o**, NAD⁺ in livers (ctrl, $n=6$; TES-991, $n=7$; TES-1025, $n=9$ mice), kidneys (ctrl, TES-1025, $n=8$; TES-991, $n=9$ mice) and brains (ctrl, $n=11$; TES-991, TES-1025, $n=12$ mice) of mice fed with normal chow diet or supplemented with ACMSD inhibitors (15 mg kg⁻¹ body weight day⁻¹). Data are mean \pm s.e.m.; each n represents a biologically independent sample. Experiments in **a–n** performed independently at least twice. * $P\leq 0.05$, ** $P\leq 0.01$, *** $P\leq 0.001$. P values calculated using two-tailed *t*-test (**a–c, f–i**, **g, j, k, m–o**), or one-way (**d**, **e**) or two-way ANOVA (**h, i**). For gel source images, see Supplementary Fig. 1. For individual P values, see Source Data.

independent of the developmental stage at which worms were exposed to *acsd-1* RNAi, but required *daf-16* (Extended Data Fig. 1r, s).

ACMSD function is conserved in mammals

We next investigated whether the molecular mechanisms observed in *C. elegans* upon *acsd-1* LOF are evolutionarily conserved. In humans,

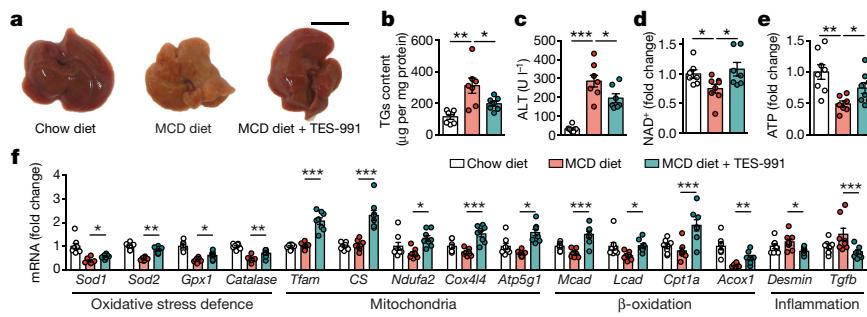


Fig. 4 | ACMSD inhibitors protect hepatic function from MCD diet-induced NAFLD. **a**, Comparison of gross liver morphology in representative 16-week-old C57BL/6J male mice fed for 2.5 weeks with control diet, MCD diet or MCD diet supplemented with 15 mg kg⁻¹ day⁻¹ of TES-991. The in vivo MCD diet study was performed once. Scale bar, 1 cm. **b–e**, Liver triglycerides (**b**), plasma ALT (**c**), liver NAD⁺ (**d**) and ATP (**e**) levels in the mouse cohorts described in **a** (control diet, MCD,

n = 8; MCD + TES-991, *n* = 7 mice). **f**, mRNA levels of oxidative stress defence, mitochondrial, β-oxidation, inflammatory and fibrosis genes in livers of the mice described in **a** (*n* = 8 mice). *Lcad* is also known as *Acadm*, *Tgfb* is also known as *Tgfb1* and *Mcad* is also known as *Acadm*. Data are mean ± s.e.m. **P* ≤ 0.05, ***P* ≤ 0.01, ****P* ≤ 0.001. *P* values calculated using two-tailed *t*-test. For individual *P* values, see Source Data.

ACMSD was mainly detected in liver and kidney. Some *Acmsd* transcripts were present in brain, 1300- and 30-fold lower than in kidney and liver, respectively²¹; this is in line with data in the Human Protein Atlas²² (<http://www.proteinatlas.org/ENSG00000153086-ACMSD/tissue>). Mouse primary hepatocytes exhibited higher levels of *Acmsd* expression than the tested mouse (AML-12, Hepa 1-6) and human (HepG2, HuH-7, HEK 293, HK-2) hepatic and renal cell lines (Extended Data Fig. 2a), and were therefore selected for further study.

Acmsd expression was reduced by more than 98% in mouse primary hepatocytes transduced with an adenovirus encoding a short hairpin RNA (shRNA) targeting *Acmsd* (Extended Data Fig. 2b). Consistent with our *C. elegans* data, total cellular NAD⁺ was 1.4-fold higher in hepatocytes treated with *Acmsd* shRNA compared to hepatocytes transduced with scrambled control shRNA (Fig. 2a), whereas there was no detectable effect on mitochondrial NAD⁺ (Extended Data Fig. 2c). *Acmsd* downregulation also enhanced mitochondrial function in hepatocytes, as indicated by increased mtDNA:nDNA ratio, citrate synthase activity and expression of OXPHOS genes both at the transcript and protein levels (Fig. 2b–e). In-gel activity assays for complex I and complex IV showed overall increased activity in cells treated with *Acmsd* shRNA, including in high molecular mass supercomplexes (Extended Data Fig. 2d, e). *Acmsd* shRNA also increased levels of UPR^{mt} transcripts (Fig. 2d) and basal and maximal oxygen consumption rate and ATP concentration in mouse primary hepatocytes grown in high or low glucose (Fig. 2f, g).

In sum, these changes indicate that hepatocyte mitochondrial function was improved by *Acmsd* LOF; we therefore tested whether knocking down *Acmsd* could render hepatocytes more resilient to steatosis and apoptosis caused by high fatty acid concentration. LOF of *Acmsd* protected primary mouse hepatocytes from fatty acid-induced apoptosis (Fig. 2h) and attenuated triglyceride accumulation (Fig. 2i); this effect was dependent on sirtuin 1 (SIRT1) (Fig. 2j, Extended Data Fig. 2f). *Acmsd* shRNA also reduced levels of ROS in primary hepatocytes in a SIRT1-dependent manner (Fig. 2k) and lowered acetylation levels of FOXO1, a deacetylation target of SIRT1 (Extended Data Fig. 2g).

Characterization of ACMSD inhibitors

The first indications that inhibiting ACMSD activity could enhance de novo NAD⁺ synthesis came from studies on phthalate esters and pyrazinamide; however, these compounds have pleiotropic effects^{23,24}. We therefore developed selective and potent ACMSD inhibitors, TES-991 and TES-1025²⁵, and investigated their therapeutic potential. Similar to genetic *Acmsd* LOF, pharmacological inhibition of ACMSD with either compound dose-dependently increased NAD⁺ levels and resulted in SIRT1 activation in primary hepatocytes (Fig. 3a–c). Both ACMSD inhibitors induced mitochondrial transcripts and enhanced both basal and maximal oxygen consumption rate and mitochondrial

SOD2 activity in primary mouse hepatocytes (Fig. 3d, e, Extended Data Fig. 3a–c). Similar to genetic *Acmsd* LOF, ACMSD inhibition protected primary hepatocytes from apoptosis induced by high doses of fatty acids (Fig. 3f) and promoted fatty acid oxidation in a SIRT1-dependent manner; this was confirmed by both induction of β-oxidation genes and functional assays (Fig. 3g–i, Extended Data Fig. 3d).

TES-1025, the inhibitor with the best exposure profile for the kidney²⁵, also increased NAD⁺ content in human HK-2 kidney cells (Fig. 3j). TES-1025 induced transcription of mitochondrial and ROS defence genes, as well as protein levels of the OXPHOS subunits in a SIRT1-dependent manner (Fig. 3k, l, Extended Data Fig. 3e). Treatment of HK-2 cells with TES-1025 also increased their ATP content (Fig. 3m). Cisplatin-induced apoptosis was attenuated when HK-2 cells were either pre-treated with (Fig. 3n) or exposed to TES-1025 simultaneously with cisplatin (Extended Data Fig. 3f).

We next characterized the effect of both compounds in the mouse. Feeding 9-week-old male C57BL/6J mice for 10 days with chow diet supplemented with TES-991 or TES-1025 (15 mg kg⁻¹ body weight day⁻¹) did not result in any pronounced effects on metabolic homeostasis and did not otherwise affect liver or kidney function (Extended Data Fig. 3g). Both TES-991 and TES-1025 increased NAD⁺ content in liver, kidneys and brain (Fig. 3o), whereas NAD⁺ levels in heart and skeletal muscle seemed unaffected (not shown). Changes in quinolinic acid did not show a consistent pattern after pharmacological inhibition of ACMSD. Whereas in brain, quinolinic acid concentration tended to increase (statistically significant for TES-1025, but to concentrations too low to provoke neurotoxicity²⁶), levels were reduced in kidney and unchanged in liver (Extended Data Fig. 3h). Both ACMSD inhibitors caused depletion of nicotinic acid in liver, kidney and brain, which could serve as another readout for ACMSD inhibition (Extended Data Fig. 3i). Besides this, administration of ACMSD inhibitors induced transcription of mitochondrial genes in the liver, whereas the expression of the same genes in the kidneys was unaffected (Extended Data Fig. 3j, k).

Efficacy of ACMSD inhibitors in disease models

On the basis of the observed differences in exposure profiles of the two ACMSD inhibitors, with TES-991 being enriched in liver and TES-1025 in the kidney²⁵, we assessed the translational potential of TES-991 in the setting of liver disease and oriented in vivo studies with TES-1025 to the kidney. Non-alcoholic fatty liver disease (NAFLD) is the most common liver disease worldwide, often leading to progressive liver damage, termed non-alcoholic steatohepatitis²⁷. Given the efficacy of boosting NAD⁺ to attenuate NAFLD in mouse models^{28,29}, we tested the efficacy of ACMSD inhibition in a mouse model of NAFLD, induced by feeding 13-week-old C57BL/6J mice with a methionine-choline-deficient (MCD) diet for 2.5 weeks²⁹. Supplementing the MCD diet with TES-991 (15 mg kg⁻¹ day⁻¹ prophylactically) attenuated hepatic

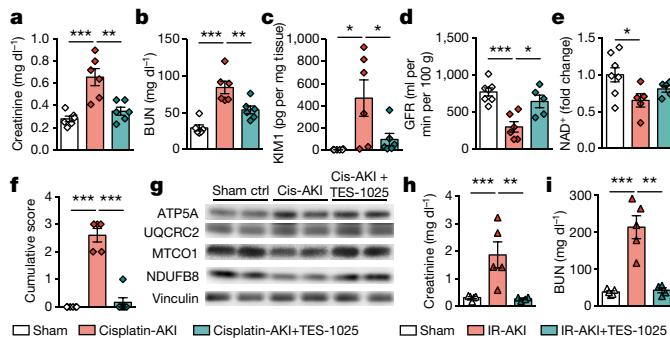


Fig. 5 | ACMSD inhibitors protect renal function in two different models of AKI. **a–c,** Plasma creatinine (**a**), BUN (**b**) and renal KIM1 protein levels (**c**) in 12-week-old C57BL/6J male mice 72 h after cisplatin administration. AKI was induced at day 10 after the beginning of the study by a single intraperitoneal injection of cisplatin (20 mg kg^{-1} body weight). Sham controls were injected with saline solution; TES-1025 was administered at 15 mg kg^{-1} body weight day $^{-1}$ ($n=6$ mice). **d,** Glomerular filtration rate (GFR) in mice 52 h after cisplatin or saline administration (sham, $n=7$; cisplatin, $n=6$; cisplatin + TES-1025, $n=5$ mice). **e,** NAD $^{+}$ levels in kidneys of the cisplatin-AKI mice (sham, $n=7$; cisplatin, cisplatin + TES-1025, $n=5$ mice). **f,** Histopathological scoring of haematoxylin and eosin (H&E)-stained kidney sections from mice described in **a**, showing tubular necrosis, tubular dilation, inflammation, oedema and cast (sham, cisplatin + TES-1025, $n=6$; cisplatin, $n=5$ mice). Scoring was performed by two pathologists in a blinded and independent way. **g,** Protein expression of respiratory complex subunits in kidneys from mouse cohorts described in **a**. The experiment was performed twice. Cis, cisplatin. **h, i,** AKI was induced in 12-week-old C57BL/6J male mice by a dorsal surgical incision and bilateral occlusion of the renal pedicles for 25 min; a simple dorsal incision was performed for the sham controls. TES-1025 was administered at 15 mg kg^{-1} day $^{-1}$. Plasma creatinine (**h**) and BUN (**i**) in mice 48 h after surgery ($n=5$ mice). Data are mean \pm s.e.m. $^*P \leq 0.05$, $^{**}P \leq 0.01$, $^{***}P \leq 0.001$. P values calculated using two-tailed t -test. For gel source images, see Supplementary Fig. 1. For individual P values, see Source Data.

steatosis (Fig. 4a, b) and plasma alanine transaminase (ALT) and aspartate transaminase levels (Fig. 4c, Extended Data Fig. 4a). As expected, the MCD diet depleted the hepatic NAD $^{+}$ pool; this effect was reversed by TES-991 (Fig. 4d). TES-991 also protected against hepatic lipid accumulation and attenuated inflammation (Extended Data Fig. 4b, c). Hepatic SOD2 activity and ATP content, which were reduced with the MCD diet, also partially recovered with TES-991 (Fig. 4e, Extended Data Fig. 4d). Furthermore, inhibition of ACMSD reversed changes in transcription of genes involved in ROS defence, β -oxidation, inflammation and mitochondrial function, which are known to be modulated in NAFLD (Fig. 4f). Whereas supplementation of the MCD diet with TES-991 was still able to increase hepatic NAD $^{+}$ in *Sirt1* knockout (*Sirt1* $^{hep-/-}$) mice (Extended Data Fig. 4e), TES-991 no longer protected the *Sirt1* $^{hep-/-}$ livers from NAFLD induced by MCD diet, as reflected in the hepatic lipid content, plasma markers of liver damage, SOD2 activity and changes in liver transcript levels (Extended Data Fig. 4f–k).

Acute kidney injury (AKI) affects 3–7% of all hospitalized patients and has a high mortality rate, especially when it occurs in the setting of intensive care 30 . A recent study found that increased NAD $^{+}$ levels can provide protection against AKI 31 . Following our observation of a protective effect of ACMSD inhibitors in cisplatin-challenged HK-2 cells, we investigated the effect of ACMSD inhibition in a mouse model of AKI, induced by application of single intraperitoneal dose of cisplatin to 12-week-old male C57BL/6J mice (Extended Data Fig. 5a). Supplementation of chow diet with TES-1025 (15 mg kg^{-1} day $^{-1}$ given prophylactically), protected against AKI, as indicated by the normalization of blood creatinine, blood urea nitrogen (BUN) and KIM1 levels (Fig. 5a–c). Glomerular filtration rate was severely compromised after cisplatin challenge in vehicle-treated animals, but was preserved in

mice receiving TES-1025 (Fig. 5d). Whereas the administration of cisplatin depleted renal NAD $^{+}$ levels, this drop in NAD $^{+}$ was less pronounced in mice treated with TES-1025 (Fig. 5e). Consistent with these biochemical indications of improved function, the increase in cumulative histopathological score—which evaluates tubular necrosis, tubular dilation, inflammation, oedema and cast formation—was also reduced by ACMSD inhibition (Fig. 5f, Extended Data Fig. 5b–d). TES-1025 also restored the cisplatin-induced changes in expression of OXPHOS complexes in the kidney (Fig. 5g).

We further confirmed the protective effects of TES-1025 (15 mg kg^{-1} body weight day $^{-1}$) in a distinct AKI model; ischaemia-reperfusion AKI (IR-AKI). Renal pedicles of 12-week-old male C57BL/6J mice were clamped for 25 min to induce mild-to-moderate AKI (Extended Data Fig. 5e). Similar to the cisplatin-induced AKI model, administration of TES-1025 protected from structural and functional renal damage inflicted by IR-AKI (Fig. 5h, i, Extended Data Fig. 5f–j). Glutathione depletion—caused by IR-AKI—was rescued and myeloperoxidase activity, reflecting neutrophil infiltration, was attenuated by TES-1025 (Extended Data Fig. 5k, l). Furthermore, increased NAD $^{+}$ levels were detected in kidneys of mice receiving TES-1025 (Extended Data Fig. 5m). TES-1025 also normalized the renal expression of OXPHOS complexes, which was impaired by ischaemia-reperfusion (Extended Data Fig. 5n).

Conclusions

Accumulating evidence supporting a key role of robust NAD $^{+}$ homeostasis in protection against ageing and numerous diseases stimulated an interest in approaches to maintain and/or increase tissue NAD $^{+}$ levels 5 . Whereas the translational potential of increasing NAD $^{+}$ produced from precursor molecules via salvage pathways has been extensively studied, little is known about the possibility of increasing NAD $^{+}$ via the de novo NAD $^{+}$ synthesis pathway. Recent reports emphasizing the role of de novo NAD $^{+}$ biosynthesis in maintenance of whole-body NAD $^{+}$ homeostasis have identified this pathway as a target to boost NAD $^{+}$ content 32,33 .

Our data position the evolutionarily conserved enzyme ACMSD as a key regulator on the de novo NAD $^{+}$ synthesis pathway. We show that ACMSD acts as a tissue-selective modulator of cellular NAD $^{+}$ levels, sirtuin activity and mitochondrial homeostasis in *C. elegans* and mouse. The high expression levels of ACMSD in the kidney and liver in mammals suggests the therapeutic potential of ACMSD inhibition for diseases of these organs such as NAFLD, non-alcoholic steatohepatitis, AKI and chronic kidney disease, which—despite their high prevalence—represent large unmet medical needs. We demonstrate beneficial effects of two potent and selective ACMSD inhibitors in animal models of NAFLD and AKI. In vivo ACMSD inhibition is associated with increases in tissue NAD $^{+}$ levels and SIRT1 activation. The ACMSD inhibitors were well-tolerated in mice and devoid of systemic side effects such as the accumulation of the potentially neurotoxic metabolite quinolinic acid 26 . In sum, our data support further investigations into the long-term benefits of ACMSD inhibition in a broader range of disease models and potential future translation of ACMSD inhibition in a clinical setting.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0645-6>.

Received: 18 September 2017; Accepted: 18 September 2018;

Published online 24 October 2018.

1. Houtkooper, R. H., Pirinen, E. & Auwerx, J. Sirtuins as regulators of metabolism and healthspan. *Nat. Rev. Mol. Cell Biol.* **13**, 225–238 (2012).
2. Imai, S. & Guarente, L. It takes two to tango: NAD $^{+}$ and sirtuins in aging/longevity control. *Nucl. Aging Mech. Dis.* **2**, <https://doi.org/10.1038/njam.2016.17> (2016).
3. Belenky, P., Bogan, K. L. & Brenner, C. NAD $^{+}$ metabolism in health and disease. *Trends Biochem. Sci.* **32**, 12–19 (2007).

4. Yang, Y. & Sauve, A. A. NAD⁺ metabolism: bioenergetics, signaling and manipulation for therapy. *Biochim. Biophys. Acta* **1864**, 1787–1800 (2016).
5. Katsyuba, E. & Auwerx, J. Modulating NAD⁺ metabolism, from bench to bedside. *EMBO J.* **36**, 2670–2683 (2017).
6. Bender, D. A. Biochemistry of tryptophan in health and disease. *Mol. Aspects Med.* **6**, 101–197 (1983).
7. Fukuoka, S. I. Identification and expression of a cDNA encoding human α -amino- β -carboxymuconate- ϵ -semialdehyde decarboxylase (ACMSD). A key enzyme for the tryptophan-niacin pathway and “quinolinate hypothesis”. *J. Biol. Chem.* **277**, 35162–35167 (2002).
8. Vrabilik, T. L., Huang, L., Lange, S. E. & Hanna-Rose, W. Nicotinamidase modulation of NAD⁺ biosynthesis and nicotinamide levels separately affect reproductive development and cell survival in *C. elegans*. *Development* **136**, 3637–3646 (2009).
9. Rongvaux, A., Andris, F., Van Gool, F. & Leo, O. Reconstructing eukaryotic NAD metabolism. *BioEssays* **25**, 683–690 (2003).
10. McReynolds, M. R., Wang, W., Holleran, L. M. & Hanna-Rose, W. Uridine monophosphate synthetase enables eukaryotic de novo NAD⁺ biosynthesis from quinolinic acid. *J. Biol. Chem.* (2017).
11. Mouchiroud, L. et al. The NAD⁺/sirtuin pathway modulates longevity through activation of mitochondrial UPR and FOXO signaling. *Cell* **154**, 430–441 (2013).
12. Hashimoto, T., Horikawa, M., Nomura, T. & Sakamoto, K. Nicotinamide adenine dinucleotide extends the lifespan of *Caenorhabditis elegans* mediated by *sir-2.1* and *daf-16*. *Biogerontology* **11**, 31–43 (2010).
13. Gebauer, J. et al. A genome-scale database and reconstruction of *Caenorhabditis elegans* metabolism. *Cell Syst.* **2**, 312–322 (2016).
14. Burnett, C. et al. Absence of effects of Sir2 overexpression on lifespan in *C. elegans* and *Drosophila*. *Nature* **477**, 482–485 (2011).
15. Viswanathan, M. & Guarente, L. Regulation of *Caenorhabditis elegans* lifespan by *sir-2.1* transgenes. *Nature* **477**, E1–E2 (2011).
16. Houtkooper, R. H. et al. Mitonuclear protein imbalance as a conserved longevity mechanism. *Nature* **497**, 451–457 (2013).
17. Benedetti, C., Haynes, C. M., Yang, Y., Harding, H. P. & Ron, D. Ubiquitin-like protein 5 positively regulates chaperone gene expression in the mitochondrial unfolded protein response. *Genetics* **174**, 229–239 (2006).
18. Haynes, C. M., Yang, Y., Blais, S. P., Neubert, T. A. & Ron, D. The matrix peptide exporter HAF-1 signals a mitochondrial UPR by activating the transcription factor ZC376.7 in *C. elegans*. *Mol. Cell* **37**, 529–540 (2010).
19. Berdichevsky, A., Viswanathan, M., Horvitz, H. R. & Guarente, L. C. *elegans* SIR-2.1 interacts with 14-3-3 proteins to activate DAF-16 and extend life span. *Cell* **125**, 1165–1177 (2006).
20. Honda, Y. & Honda, S. The *daf-2* gene network for longevity regulates oxidative stress resistance and Mn-superoxide dismutase gene expression in *Caenorhabditis elegans*. *FASEB J.* **13**, 1385–1393 (1999).
21. Pucci, L., Perozzi, S., Cimadamore, F., Orsomando, G. & Raffaelli, N. Tissue expression and biochemical characterization of human 2-amino 3-carboxymuconate 6-semialdehyde decarboxylase, a key enzyme in tryptophan catabolism. *FEBS J.* **274**, 827–840 (2007).
22. Uhlén, M. et al. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
23. Fukuyatari, T. Phthalate esters enhance quinolinate production by inhibiting α -amino- β -carboxymuconate- ϵ -semialdehyde decarboxylase (ACMSD), a key enzyme of the tryptophan pathway. *Toxicol. Sci.* **81**, 302–308 (2004).
24. Saito, K. et al. Mechanism of increases in L-kyurenine and quinolinic acid in renal insufficiency. *Am. J. Physiol. Renal Physiol.* **279**, F565–F572 (2000).
25. Pellicciari, R. et al. α -amino- β -carboxymuconate- ϵ -semialdehyde decarboxylase (ACMSD) inhibitors as novel modulators of de novo nicotinamide adenine dinucleotide (NAD⁺) biosynthesis. *J. Med. Chem.* **61**, 745–759 (2018).
26. Chen, Y. & Guillemin, G. J. Kynurene pathway metabolites in humans: disease and healthy states. *Int. J. Tryptophan Res.* **2**, 1–19 (2009).
27. Michelotti, G. A., Machado, M. V. & Diehl, A. M. NAFLD, NASH and liver cancer. *Nat. Rev. Gastroenterol. Hepatol.* **10**, 656–665 (2013).
28. Gariani, K. et al. Eliciting the mitochondrial unfolded protein response by nicotinamide adenine dinucleotide repletion reverses fatty liver disease in mice. *Hepatology* **63**, 1190–1204 (2016).
29. Gariani, K. et al. Inhibiting poly-ADP ribosylation increases fatty acid oxidation and protects against fatty liver disease. *J. Hepatol.* (2016).
30. Lewington, A. J., Cerdá, J. & Mehta, R. L. Raising awareness of acute kidney injury: a global perspective of a silent killer. *Kidney Int.* **84**, 457–467 (2013).
31. Tran, M. T. et al. PGC1α drives NAD biosynthesis linking oxidative metabolism to renal protection. *Nature* **531**, 528–532 (2016).
32. Liu, L. et al. Quantitative analysis of NAD synthesis-breakdown fluxes. *Cell Metab.* **27**, 1067–1080 (2018).
33. Shi, H. et al. NAD deficiency, congenital malformations, and niacin supplementation. *N. Engl. J. Med.* **377**, 544–552 (2017).

Acknowledgements We thank P. Gönczy and the *Caenorhabditis* Genetics Center for providing reagents, the Bioimaging and Optics Core Facility and the Phenotyping Unit of EPFL, N. Moullan, T. Clerc and S. Bichet for technical assistance. E.K. was supported by Fondation Romande pour la Recherche sur le Diabète. M.Z. was supported by the KNOW consortium ‘Healthy Animal—Safe Food’ MS&HE no. 05-1/KNOW2/2015 and the Foundation for Polish Science. This work was supported by funds from EPFL and Swiss National Science Foundation (grant 3100030B_160318). J.I. acknowledges funding from the Foundation Pierre-Mercier pour la Science. We thank H. Gallant-Ayala for advice on analytical methods.

Reviewer information *Nature* thanks W. Hanna-Rose, S. Parikh, J. Tam and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions E.K., J.A. and R.P. conceived and designed the project. E.K., A.M., M.Z., F.D.F., K.G., D.R. and N.G. performed experiments. P.L. synthesized ACMSD inhibitors. O.M. made the *acsd-1::GFP* *C. elegans* reporter strain. N.R. and L.C. performed activity assays for the enzymes in de novo NAD⁺ biosynthesis. V.v.d.V. and J.I. performed the metabolomics and N.S.-R. carried out histopathology. S.d.S. and D.L. assisted with GFR measurements. E.K., K.S. and J.A. wrote the manuscript with the contributions from all other authors.

Competing interests J.A., R.P. and N.R. are inventors on US patent 9,708,272 (18 July, 2017), filed by TES Pharma S.r.l., Corciano, Italy. The patent covers the results obtained with the ACMSD inhibitors, TES-991 and TES-1025, described in Figs. 3–5. R.P., F.D.F., N.G. and P.L. are employed by TES Pharma.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0645-6>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0645-6>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to R.P. or J.A.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

C. elegans experiments. *C. elegans* strains were provided by the *Caenorhabditis* Genetics Center (University of Minnesota). Worms were maintained on nematode growth medium (NGM) agar plates seeded with *E. coli* OP50 at 20 °C, unless stated otherwise. The strains used for the experiments are the following: Bristol N2, NL2099 (*rrf-3(pk1426)II*), SJ4143 (*zcls17[ges-1::GFP(mit)]*), SJ4103 (*zcls14[myo-3::GFP(mit)]*), KN259 (*hul33[sod-3::GFP+pRF4(rol-6(su1006))]*), SJ4005 (*zcls4[hsp-4::GFP]*), TJ356 (*zls356[daf-16p::daf-16a/b::GFP+rol-6(su1006)]*), SJ4100 (*zcls13[hsp-6::GFP]*), VC199 (*sir-2.1(ok434)IV*). If not indicated explicitly in the text, the strain used for experiments was NL2099 (*rrf-3(pk1426)II*).

Bacterial feeding RNAi experiments were carried out as previously described³⁴. Clones used were *acs-1* (Y71D11A.3), *sir-2.1* (R11A8.4), *daf-16* (R13H8.1), *ubl-5* (F46F11.4) and *atfs-1* (ZC376.7). Clones were purchased from GeneService and their identity was confirmed by sequencing. For the double-RNAi experiments, bacterial cultures were mixed before seeding on NGM plates. The control RNAi in these experiments was diluted 50% with control empty vector RNAi bacteria. **Lifespan assays.** *C. elegans* lifespan assays were carried at 20 °C as previously described³⁵. One hundred worms were used per condition and scored every 2 days. Reasons for censoring were the exploded vulva phenotype or worms that crawled off the plate and were pre-established before the beginning of the experiment. Where indicated, paraquat dichloride (Sigma-Aldrich) was added on top of the agar plates to obtain the indicated final concentration. Once the paraquat solution was absorbed completely by the agar, L4 worms were transferred to these agar plates and monitored daily for 5–6 days. Tests were performed in non-blinded manner and repeated twice, unless indicated otherwise.

Mobility assessment. The movement of worms was recorded for 45 s at days 1, 3 and 5 of adulthood using a Nikon DS-L2/DS-Fi1 camera and controller setup, attached to both a computer and a standard bright-field microscope. For each condition, 5 plates of worms, with 10 worms per plate, were used. The movement of worms was calculated by taking an integral of the speed value, which was assessed by following the worm centroids with a modified version of the freely-available 'parallel worm tracker' for MATLAB³⁶. The experiments were non-blinded and repeated twice.

Generation of transgenic *C. elegans* strains. *acs-1p::GFP* expression vector was created by amplifying the 1,425-bp sequence upstream from the transcription start site of the *acs-1* coding region by using *C. elegans* genomic DNA and 5'-gtaACATGTcagtcgacgcaaaaattgtt-3' and 5'-gtaCCCGGTTtgcaggaaaattataaaaattaaatg-3' primers. The PCR product was digested with PciI and XmaI and ligated into the PciI and AgeI-digested *pPD30.38* expression vector containing the *gfp* coding sequence cloned between inserted AgeI and NotI restriction sites. Two independent lines carrying *acs-1p::GFP* transgene as extrachromosomal array were analysed in the study.

Imaging and image processing. Confocal images were acquired with Zeiss LSM 700 Upright confocal microscope (Carl Zeiss AG) under non-saturating exposure conditions. The worms were prepared for imaging as previously described¹⁶. In brief, nematodes were immobilized with 7.5 mM solution of tetramisole hydrochloride (Sigma-Aldrich) in M9 and mounted on 6% agarose pads on glass slides. Image processing was performed with Fiji software (<http://imagej.nih.gov/ij>; v.1.47b). Each experiment was repeated at least once.

GFP quantification. Fluorescence intensity in worm strains expressing GFP-reporter proteins was quantified using a Victor X4 plate reader (Perkin Elmer). The animals were prepared in the following way: 80 worms per condition (at the corresponding ages) were transferred into wells filled with M9 medium (20 worms per well of a black-walled 96-well plate). Each experiment was repeated at least twice. **MitoSox ROS quantification.** MitoSox staining was performed as previously described¹¹. In brief, a population of 100 worms was recovered in 1 ml M9 buffer, washed five times to remove residual bacteria, and resuspended in 200 µl 1:200 MitoSox (Life technologies) stock solution (initial stock solution was dissolved at 1 mM in DMSO). After 20 min of treatment, worms were washed five times in 1 ml M9 buffer to eliminate the MitoSox reagent, and then transferred to a black-walled 96-well plate for reading using a Victor X4 plate reader (Perkin Elmer).

Enzymatic assays in *C. elegans*. Worms were collected in M9 and washed eight times to remove all residual bacteria. After the last wash, the maximum possible amount of M9 was aspirated and the remaining worm pellet was snap-frozen in liquid nitrogen. Activity of QPRT was assessed as previously described³⁷. In brief, we coupled the QPRT reaction to the conversion of the reaction product NAMN to NAD⁺ with the help of the ancillary enzymes NadD and NadE, followed by a quantification of NAD⁺ with fluorometric cycling assay (Extended Data Fig. 1c). The buffer of the assay mixture consisted of 30 mM potassium phosphate buffer, pH 7.0, 0.5 mM MgCl₂. ACMSD activity was assayed as previously described²¹. In brief, formation of ACMS was monitored at 360 nm, at 37 °C, in a pre-assay mixture consisting of 30 µM hydroxyanthranilic acid and an excess quantity of recombinant *Ralstonia metallidurans* 3-hydroxyanthranilic acid dioxygenase. After the reaction was complete, an appropriate aliquot of worm extract was added and ACMS

consumption was calculated by the decrease in absorbance. The activity value was corrected for the spontaneous decrease in absorbance due to the non-enzymatic cyclization of ACMS to quinolinic acid. To this end, a control mixture containing all reagents except the extract (replaced by an equal volume of a BSA solution at the same concentration of the extract) was monitored in parallel.

Tryptophan supplementation experiments in *C. elegans*. L-tryptophan (Sigma-Aldrich) was dissolved in water to obtain a 50 mM solution. This freshly prepared solution was added into the NGM agar at the moment of plate preparation to obtain the concentrations stated in the main text. Worms were maintained on the tryptophan-supplemented plates seeded with live bacteria (*E. coli* OP50) through their larval stages to allow development. At L4 larvae stage, worms were collected in M9 solution, washed seven times to remove all the residual bacteria in their intestines and transferred to tryptophan-supplemented plates seeded with heat-inactivated bacteria (*E. coli* OP50) to avoid metabolism of L-tryptophan by the bacteria. A 500 ml culture of OP50 bacteria was incubated at 37 °C overnight; the bacteria were then centrifuged at 3,320g for 30 min at 4 °C and the resulting bacterial pellet was dissolved in 50 ml fresh LB medium. Subsequently, heat inactivation was performed by incubating the bacteria at 95 °C for 30 min.

NAD⁺ measurements. NAD⁺ was extracted using acidic extraction method and analysed by HPLC-mass spectrometry as previously described³⁸. In brief, ~10 mg of frozen tissue samples were used for NAD⁺ extraction in 10% perchloric acid and neutralized in 3 M K₂CO₃ on ice. After final centrifugation, the supernatant was filtered and the internal standard (NAD⁺-C13) was added and loaded onto a column (150 Å, ~2.1 mm; Kinetex EVO C18, 100 Å). HPLC was run for 1 min at a flow rate of 300 ml/min with 100% buffer A (methanol/H₂O, 80/20% v/v). Then, a linear gradient to 100% buffer B (H₂O + 5 mM ammonium acetate) was performed (at 1 to 6 min). Buffer B (100%) was maintained for 3 min (at 6 to 9 min), and then a linear gradient back to 100% buffer A (at 9 to 13 min) started. Buffer A was then maintained at 100% until the end (at 13 to 18 min). NAD⁺ eluted as a sharp peak at 3.3 min and was quantified on the basis of the peak area ratio between NAD⁺ and the internal standard and normalized to tissue weight and protein content.

Cell culture. The mouse hepatocytes cell line AML-12 (alpha mouse liver 12) was obtained from ATCC and grown at 37 °C in a humidified atmosphere of 5% CO₂/95% air in Dulbecco's Modified Eagle Medium/Nutrient Mixture F-12 (DMEM/F-12) supplemented with 0.005 mg/ml insulin, 0.005 mg/ml transferrin, 5 ng/ml selenium, 40 ng/ml dexamethasone, 0.5 mM tryptophan and 1% gentamycin.

Proximal tubular cell line HK-2 (human kidney 2) was purchased from ATCC and grown at 37 °C in a humidified atmosphere of 5% CO₂/95% air in normal DMEM medium (Gibco) including 10% FCS (Gibco), 10 units per ml penicillin, 0.5 mM tryptophan and HEPES for buffering. ACMSD inhibitor was initially diluted from powder in DMSO to the stock concentration of 1 mM. This stock was further diluted with water to 100 µM, which was used for the cell treatments.

All the cell lines purchased from ATCC have been authenticated by morphology, karyotyping and PCR-based approaches. All the used cell lines have been routinely checked in the laboratory for mycoplasma contamination with the MycoProbe detection kit (R&D systems). Only cells negative for mycoplasma contamination were used.

Primary hepatocytes culture. Primary hepatocytes were prepared from 8 to 12-week-old C57BL/6 or *Sirt1*^{L2/L2} mice (males and females) by liberase perfusion method as previously described³⁹, with minor modifications. Isolated primary hepatocytes were plated with DMEM medium (Gibco) including 10% FCS (Gibco), 10 units per ml penicillin, 0.5 mM tryptophan and HEPES for buffering. After 6–8 h of attachment, this medium was replaced with medium containing either different concentrations of ACMSD inhibitor or the corresponding concentration of DMSO, or transduced with adenovirus encoding either control shRNA or shRNA against *Acmsd*. Primary hepatocytes extracted from *Sirt1*^{L2/L2} mice were transduced either with adenovirus encoding CRE-recombinase (to generate *Sirt1* knockout) or GFP. Primary hepatocytes were collected 24 h later in the case of pharmacological treatment and 48 h after adenoviral transduction. When indicated, cell culture medium was supplemented with palmitate-BSA and/or oleate-BSA.

Genetic knockdown of *Acmsd*. Five different pairs of single-stranded DNA oligonucleotides were designed and tested for *Acmsd* knockdown. 'Top strand' oligonucleotides: 1) 5'-caccggagacttcagagtatccgaaggatcactctgaagacttc-3', 2) 5'-caccggagatggaggcgtgttgcataatcaacaaacggctccatctcc-3', 3) 5'-caccgtatggcagatgttgcataatcaacaaacggctccatctcc-3', 4) 5'-caccggaaactatgttgcataatcaacaaacggctccatctcc-3', 5) 5'-caccggcgttgcataatcaacaaacggctccatctcc-3'. 'Bottom strand' oligonucleotides (respectively): 1) 5'-aaaaggagatggaggcgtgttgcataatcaacaaacggctccatctcc-3', 2) 5'-aaaaggagatggaggcgtgttgcataatcaacaaacggctccatctcc-3', 3) 5'-aaaaggagatggaggcgtgttgcataatcaacaaacggctccatctcc-3', 4) 5'-aaaaggagatggaggcgtgttgcataatcaacaaacggctccatctcc-3', 5) 5'-aaaaggagatggaggcgtgttgcataatcaacaaacggctccatctcc-3'. BLOCK-iT U6 RNAi Entry Vector Kit (Invitrogen) was used for production of an entry clone. BLOCK-iT Adenoviral

RNAi Expression System (Invitrogen) was used subsequently to produce an adenoviral expression clone. The shRNA no. 3 showed the highest knockdown efficacy and was used for all the experiments in this manuscript.

Apoptosis assessment. Caspase 3/7 activities were measured with the Caspase-Glo assay (Promega) according to the manufacturer's instructions. HK-2 cells were plated at 5×10^3 cells/well in a white ViewPlate-384 microplate (PerkinElmer). The cells were treated with 50 μM cisplatin (Sigma Aldrich) for 16 h. ACMSD inhibitor was either added simultaneously with cisplatin or 1 h before cisplatin addition (preventive). The data were graphed considering the cisplatin treatment as 100% and the vehicle treatment as 0% of caspase activity. Apoptosis in primary hepatocytes was induced by exposing the cells to 0.75 mM palmitate for 36 h.

SOD2 activity assay. Primary hepatocytes, AML-12 cells or pieces of frozen liver were lysed in 20 mM HEPES buffer (Gibco), pH 7.2, containing 1 mM EGTA (Sigma-Aldrich), 210 mM mannitol (Sigma-Aldrich) and 70 mM sucrose (AMRESCO). Total protein concentration was determined using the Bradford assay (BioRad). SOD2 activity was determined at indicated times after ACMSD inhibitor treatment by the SOD Assay Kit (Cayman Chemical) according to the manufacturer's instructions. To specifically detect the SOD2 activity 2 mM potassium cyanide was added to the assay, which inhibited both Cu/Zn-SOD and extracellular SOD, resulting in the detection of only Mn-SOD (SOD2) activity. Absorbance was determined with a Victor X4 multilabel plate reader (PerkinElmer) at 450 nm. Results are expressed in U/ml/mg of protein according to the standard curve and measured protein concentration.

Oxygen consumption. Oxygen consumption was measured with the Seahorse XF96 instrument (Seahorse Bioscience) according to the manufacturer's protocol. FCCP at the indicated concentrations was used as an uncoupler to reach maximal respiration.

Fatty acid oxidation. Fatty acid oxidation was measured in primary mouse hepatocytes with the Seahorse XF96 instrument (Agilent Seahorse), following the manufacturer's protocol. This protocol quantified the oxidation of both exogenous and endogenous fatty acids. Cells were cultured in conditions that would stimulate the depletion of endogenous substrates to prime the cells for oxidation of exogenous fatty acids. Four different conditions were used for each of the treatment groups. These conditions included etomoxir+ palmitate-, etomoxir+ palmitate+, etomoxir- palmitate- and etomoxir- palmitate+. To determine endogenous fatty acid utilization etomoxir+ conditions were used, as cells treated with etomoxir were unable to import exogenous fatty acids into the mitochondria.

Steatosis assessment in cells. Extent of steatosis and triglyceride accumulation within cells were assessed with Steatosis Colorimetric Assay Kit (Cayman) and Triglyceride Colorimetric Assay Kit (Cayman), according to manufacturer's instructions. Triglycerides were normalized by protein content, which was determined using the Bradford assay (BioRad).

ROS quantification in cells. ROS levels were measured by staining cells with 20 μM 2',7'-dichlorofluorescin (DCFDA) reagent (Abcam) for 45–60 min at 37°C. The DCFDA solution was then removed and cells were washed twice with PBS. The quantification of the signal was performed with a Victor X4 multilabel plate reader (Perkin-Elmer) with maximum excitation and emission spectra of 495 nm and 529 nm, respectively.

RNA isolation and RT-PCR. Total RNA was extracted using TRIzol (Invitrogen), according to the manufacturer's instructions. RNA was treated with DNase, and 2 μg RNA was used for reverse transcription. cDNA diluted 15 \times or 20 \times was used for quantitative PCR with reverse transcription (RT-qPCR) reactions. The RT-qPCR reactions were performed using the Light-Cycler system (Roche Applied Science) and a qPCR Supermix (QIAGEN) with the indicated primers. The average of at least three technical repeats was used for each biological data point. The list of primers used and their sequences are available upon request. The experiments were repeated at least twice (starting from RNA isolation). In case of contradictions in the results between the two repetitions, the experiment was repeated for a third time.

Protein isolation and western blot. Proteins were extracted with RIPA buffer containing protease and phosphatase inhibitors and analysed by SDS-PAGE and western blot. Proteins were detected using the following specific antibodies: ACMSD (Abcam, ab96081), actin (Sigma, A5441), acetyl-FKHR (Santa Cruz, sc-49437), FOXO1 (Cell Signaling, 2880), HSP90 (BD Transduction Laboratories, 610418), Total OXPHOS Rodent Antibody Cocktail (Abcam, ab110413), tubulin (Santa Cruz, sc-5286), vinculin (Abcam, ab129002). All antibodies were validated by the manufacturer.

Mitochondrial isolation, blue native polyacrylamide gel electrophoresis and in-gel activity assays. Mitochondrial isolation was performed as previously described⁴⁰. In brief, ~30 mg of liver tissue or a pool of 10,000 worms was used for mitochondrial isolation. The samples were homogenized in ice-cold sucrose-containing isolation buffer with a glass/teflon Potter-Elvehjem homogenizer (Wheaton, cat. No. 358029) and mitochondria were pelleted through multiple rounds of centrifugation at different speeds. For blue native PAGE, 50 μg of

mitochondria from liver was solubilized in digitonin and sample buffer (Invitrogen, BN2008). Electrophoresis was performed using Native PAGE Novex Bis-Tris Gel System (3 to 12%), as per manufacturer's instructions with minor modifications. Gel transfer was performed using Invitrogen iBlot gel transfer system. For detection of the complexes, anti-OXPHOS cocktail (Invitrogen, 457999) and WesternBreeze Chromogenic Western Blot Immunodetection Kit (Invitrogen, WB7103) were used. In the final detection step, the membrane was incubated with the chromogenic substrate for 8 min for all the gels.

For in-gel activity assays, electrophoresis was performed for 3 h (30 min at 150 V and 2.5 h at 250 V). Complex I activity was performed by incubating the gels for 15 to 30 min in the substrate composed of 2 mM tris-HCl pH 7.4; 0.1 mg/ml NADH, and 2.5 mg/ml nitrotetrazolium blue. Complex IV activity was performed by incubating the gels for 30 to 40 min in the substrate composed of 25 mg of 3,3'-diamidobenzidine tetrahydrochloride; 50 mg cytochrome *c*; 45 ml 50 mM phosphate buffer pH 7.4, and 5 ml water. Complex IV + complex I activity was performed by subsequently incubating the gels in the substrate for CIV followed by incubation in complex I substrate. Complex II activity was checked on a separate gel, which was incubated in the substrate composed of 20 mM sodium succinate, 2.5 mg/ml nitrotetrazolium blue, 0.2 mM phenazine methosulfate and 5 mM tris-HCl buffer. All reactions were stopped with 10% acetic acid.

SIRT1 activity. SIRT1 activity was assessed with SIRT1 Direct Fluorescent Screening Assay Kit (Cayman Chemical), according to the manufacturer's instructions. The absorbance was read at 450 nm.

mtDNA copy number. Mitochondrial number was assessed by using mtDNA: nDNA ratio. mtDNA was quantified as previously described⁴¹, with slight modifications. In brief, total DNA was extracted with the Nucleospin Tissue Kit (Macherey Nagel). Forty nanograms total DNA was assessed by real-time PCR using a Light Cycler 480 (Roche). The reaction was performed in a final volume of 8 μl with 1 \times SYBR green master mix (Roche) and 1.25 μM of the reverse and forward primers. UCP2 primers were used as endogenous control for nDNA and 16S as marker for mtDNA.

Statistics. No statistical methods were used to predetermine sample size. Survival analyses were performed using the Kaplan-Meier method and the significance of differences between survival curves calculated using the log-rank test. Differences between two groups were assessed using two-tailed *t*-tests. To compare the interaction between age and genotype, two-way ANOVA tests were performed. Analysis of variance, assessed by Tukey's or Dunnett's multiple comparison test, was used when comparing more than two groups. We used GraphPad Prism 6 (GraphPad Software) for all other statistical analyses. Sample sizes were chosen without performing statistical tests, but based on studies with similar experimental design and on the known variability of the assay. All *P* values <0.05 were considered significant. **P* < 0.05 ; ***P* ≤ 0.01 ; ****P* ≤ 0.001 unless stated otherwise.

Mouse experiments. Mice were maintained in a controlled environment with 22°C temperature, 50% humidity, 15–20 fresh air changes per hour and light-dark cycle of 12 h. They were housed 3–4 animals per cage with ad libitum access to fresh water and food and with appropriate environmental enrichment within the cage. All animal experiments were performed according to Swiss ethical guidelines and approved by the Service de la Consommation et des Affaires Vétérinaires (SCAV) of the Canton de Vaud. Animals that showed signs of severity, predefined by the animal authorizations, were euthanized and removed from the calculations. All experiments on mice were done in a non-blinded way.

Generation of *Sirt1*^{hep-/-} mice. Liver-specific *Sirt1* knockout (*Sirt1*^{hep-/-}) mice were generated as previously described²⁹, by breeding *Sirt1*^{1/2/1/2} mice¹¹ with Alb-Cre mice, which express Cre recombinase under the control of the albumin promoter.

MCD diet-induced NAFLD. Twelve-week-old, C57BL/6J male mice were acclimatized for a period of seven days before initiation of the experiment. After the acclimatization period the mice were randomized based on their body weight and separated into three groups: one group received methionine-choline deficient (MCD) diet (Harlan Teklad TD.90262) containing vehicle (DMSO 5% + polyethylene glycol (PEG) 20%), the second received MCD diet supplemented with 15 mg kg⁻¹ body weight day⁻¹ ACMSD inhibitor (TES-991) and the third receiving the matched control diet (Harlan Teklad TF.94149) with the vehicle (DMSO 5% + PEG 20%). Two and a half weeks after the beginning of the special diets, animals fasted for 4 h were subjected to isoflurane anaesthesia and blood was sampled via cardiac puncture. Plasma was collected by centrifugation at 4,000 r.p.m. for 10 min at 4°C. Collected tissues and plasma were snap-frozen in liquid nitrogen. The experiment was performed once.

Histopathology on MCD diet liver tissues. A piece of liver tissue (always from the same lobe) was fixed with 10% neutral buffered formaldehyde, processed and embedded in paraffin. After sectioning, tissue was stained with H&E. Another liver piece (always from the same lobe) was embedded in Shandon Cryomatrix resin (Thermo Scientific Scientific) and snap-frozen in isopentane cooled in

liquid nitrogen, before being placed on dry ice. Oil red O and CD45 stainings were performed on cryosections as previously described²⁹.

Hepatic triglyceride content measurement. Hepatic lipids were extracted as previously described⁴². Triglyceride content in hepatic lipid fraction was quantified with enzymatic assays (Roche).

Ischaemia-reperfusion-induced AKI. Nine-week-old C57BL/6J males were acclimatized for seven days before initiation of the experiment. After the acclimatization period, the mice were randomized based on their body weight and separated into three groups: two groups received normal chow diet (Harlan Teklad 2916) containing vehicle (DMSO 5% + polyethylene glycol (PEG) 20%), the third group received normal chow diet supplemented with 15 mg kg⁻¹ body weight day⁻¹ ACMSD inhibitor (TES-1025). On day 10 after the beginning of the diet, mice were anaesthetized with isoflurane and placed on a surgical platform in a dorsal position. Both kidneys were exposed through flank incisions and renal pedicles were occluded using vascular clamps for 25 min. The clamp was then removed and the surgical site was sutured. Physiological saline (1 ml) was administered intraperitoneally after closing the wound to prevent dehydration. The sham-operated group was subjected to similar surgical procedures, except that the occluding clamp was not applied. A suitable analgesic (Dafalgan) was administered post-operatively to all animals. Animals were monitored for recovery from anaesthesia and then housed singly in their home cage with appropriate environmental enrichment. Forty-eight hours after the surgery, animals were subjected to isoflurane anaesthesia and blood was sampled by cardiac puncture, followed by organ collection. Plasma was collected by centrifugation at 4,000 r.p.m. for 10 min at 4°C. Plasma was separated into a fresh tube and stored at -80°C. Collected tissues were snap-frozen in liquid nitrogen. The experiment was performed once.

Cisplatin-induced AKI. Eight-week-old C57BL/6J males were acclimatized for 5–7 days before initiation of the experiment. Animals were randomized into different treatment groups based on their body weight and BUN levels. After randomization the groups were maintained on specified diet, either normal chow diet (Teklad global 16% protein rodent diet; Harlan Laboratories 2016S) or chow diet supplemented with ACMSD inhibitor, for 10 days. Body weight and food consumption were monitored during this period. Cisplatin was injected intraperitoneally with 1 ml syringe with 26G needle at a dose of 20 mg/kg of body weight. The sham control group was administered with the same volume of 0.9% saline. Animals were monitored every day for body-weight loss, general health conditions and signs of pain, distress and mortality. Seventy-two hours after cisplatin injection animals were subjected to isoflurane anaesthesia and blood was sampled by cardiac puncture, followed by organ collection. Plasma was collected by centrifugation at 4,000 r.p.m. for 10 min at 4°C. Plasma was separated into a fresh tube and stored at -80°C. Collected tissues were snap-frozen in liquid nitrogen. The experiment was performed twice.

Glomerular filtration rate measurement. Determination of glomerular function rate was performed in conscious mice using transcutaneous measurement of FITC-sinistrin (Mannheim Pharma and Diagnostic) elimination as previously described^{43,44}, using miniaturized devices containing an optical component and a microprocessor (Medibeacon). In brief, 55 h post-cisplatin administration, mice were subjected to a light (1.5–2% isoflurane) anaesthesia to affix the device and a rechargeable battery on a depilated region of the back using a double-sided sticky patch. The device was further held in place with a piece of Transpore tape (3M) surrounding the animal body. After a resting period of 2–5 min, a bolus of FITC-sinistrin (35 mg per 100 g, dissolved in saline) was injected through the tail vein. The excretion kinetics of FITC-sinistrin was recorded in conscious animals for an average of 60 min as a decay of subcutaneous fluorescence of FITC-sinistrin. Elimination half-life ($t_{1/2}$) was determined from the single exponential phase of the excretion curve and then converted to GFR using a semi-empirical conversion factor developed and validated for mice as previously described⁴⁴.

Histopathology on AKI kidneys. Half of a kidney was fixed with 10% neutral-buffered formaldehyde, processed and embedded in paraffin. After sectioning, tissue was stained with H&E. Histopathological scoring was performed by two different pathologists in a blinded and independent way. The scoring system was adapted from ref.⁴⁵ with slight modifications. Tubular cell necrosis, tubular dilatation, inflammatory cell infiltration, oedema and cast formation were scored.

Quantification of intermediates in tryptophan pathway. Snap-frozen brain, liver and kidney samples were ground into powder with liquid nitrogen, using a mortar and pestle. Each sample was then pre-weighed (~50 mg) into lysis tubes (soft tissue homogenizing CK 14 tubes) and stored at -80°C before metabolite extraction. Frozen tissue powders were extracted by the addition of an ice-cold methanol-H₂O 4:1 mixture (100 µl for every 10 mg of tissue) and homogenized using ceramic beads in the Cryolys Precellys 24-sample Homogenizer (2 × 20 s at 10,000 r.p.m.; Bertin Technologies), which was air-cooled by a flow rate at 110 l/min at 6 bar. Homogenized extracts were centrifuged for 15 min at 4,000g at 4°C and the supernatant (tissue extract) was removed and used in further preparation for liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis.

For the analysis of tryptophan pathway intermediates and NAD⁺, in negative ionization mode, the sample was prepared by mixing an aliquot of tissue extract (50 µl) with 250 µl of the ice-cold internal standard solution (in 100% acetonitrile). For the analysis of end products of kynurenine pathway (that is, picolinic, quinolinic and nicotinic acid), in positive ionization mode, 300 µl of the tissue extract was evaporated to dryness (using a CentriVap vacuum concentrator, LabConco), and reconstituted with 75 µl 0.2% formic acid in H₂O. Samples were centrifuged and supernatant was injected for LC-MS/MS analysis.

LC-MS/MS analyses were performed using a 6495 triple quadrupole mass spectrometer (QqQ) interfaced with a 1290 UHPLC system (Agilent Technologies) and operated in the dynamic Multiple Reaction Monitoring (dMRM) mode.

Quinolinic, picolinic and nicotinic acid were measured in positive electrospray ionization (ESI + MS) mode, using an Acquity HSS T3 column (2.1 × 100 mm, 1.8 µm, Waters). An 11-min gradient was applied starting at 0% B (0.2% formic acid in methanol, 0–2 min) and increasing to 50% (2–4 min), and further to 90% (4–5 min), before returning to starting conditions (5–7 min) and re-equilibrating for 4 min (7–11 min). Mobile phase A was 0.2% formic acid in H₂O, the flow rate for this method was 400 µl/min and the sample injection volume was 5 µl. ESI source conditions were set as follows: dry gas temperature 250°C, nebulizer 35 psi and flow 15 l/min, sheath gas temperature 250°C and flow 8 l/min, nozzle voltage 1,000 V, and capillary voltage +3,000 V. The cycle time in dMRM mode was 500 ms. Standard calibration curves ranging from 0–50 µM were used for quantification for all three organic acids.

NAD⁺ was measured in negative electrospray ionization (ESI-MS) mode, using a SeQuant ZIC-pHILIC column (10 × 2.1 mm internal diameter, 5 µm) with a SeQuant ZIC-pHILIC guard column (20 × 2.1 mm internal diameter, 5 µm (Merck)) at an operation temperature of 30°C. Mobile phase A was composed of 20 mM ammonium acetate and 20 mM ammonium hydroxide in H₂O (pH 9.3) and mobile phase B was 100% acetonitrile, and the sample injection volume was 2 µl. The linear gradient elution started at 90% B (0–1.5 min), decreased to 50% B (8–11 min), decreased further to 45% B (12–15 min), restored to 90% (15–16 min) and subsequently re-equilibrated for 9 min at a flow rate of 300 µl/min. ESI source conditions were set as follows: dry gas temperature 290°C, nebulizer 35 psi and flow 14 l/min, sheath gas temperature 350°C and flow 12 l/min, nozzle voltage 0 V, and capillary voltage -2,000 V. Cycle time in dMRM mode was 600 ms. The following calibration curve range was used for NAD⁺ quantification: 0–316 µM. NAD⁺ concentrations were reported with an external calibration curve only (IS not available). Collision energies and product ions (MS2 or quantifier and qualifier ion transitions) were pre-optimized for each metabolite of interest.

Data processing, including the peak integration and concentration calculation was performed in MassHunter Quantitative Analysis (for QqQ, v.B.07.01/Build 7.1.524.0, Agilent Technologies). Linearity of the standard curves was evaluated for each metabolite using a 6-point range; in addition, peak area integration was manually curated and corrected where necessary. Concentrations of the compounds for which the IS were available were corrected for the ratio of MS response (peak area) between the analyte and the IS, to account for matrix effects.

Glutathione assay. Glutathione quantities in the kidney were assayed using the ELISA kit (Cayman Chemicals) according to the manufacturer instructions.

Myeloperoxidase assay. Myeloperoxidase (MPO) in kidney tissue was quantified by ELISA kit (Abcam) according to the manufacturer instructions.

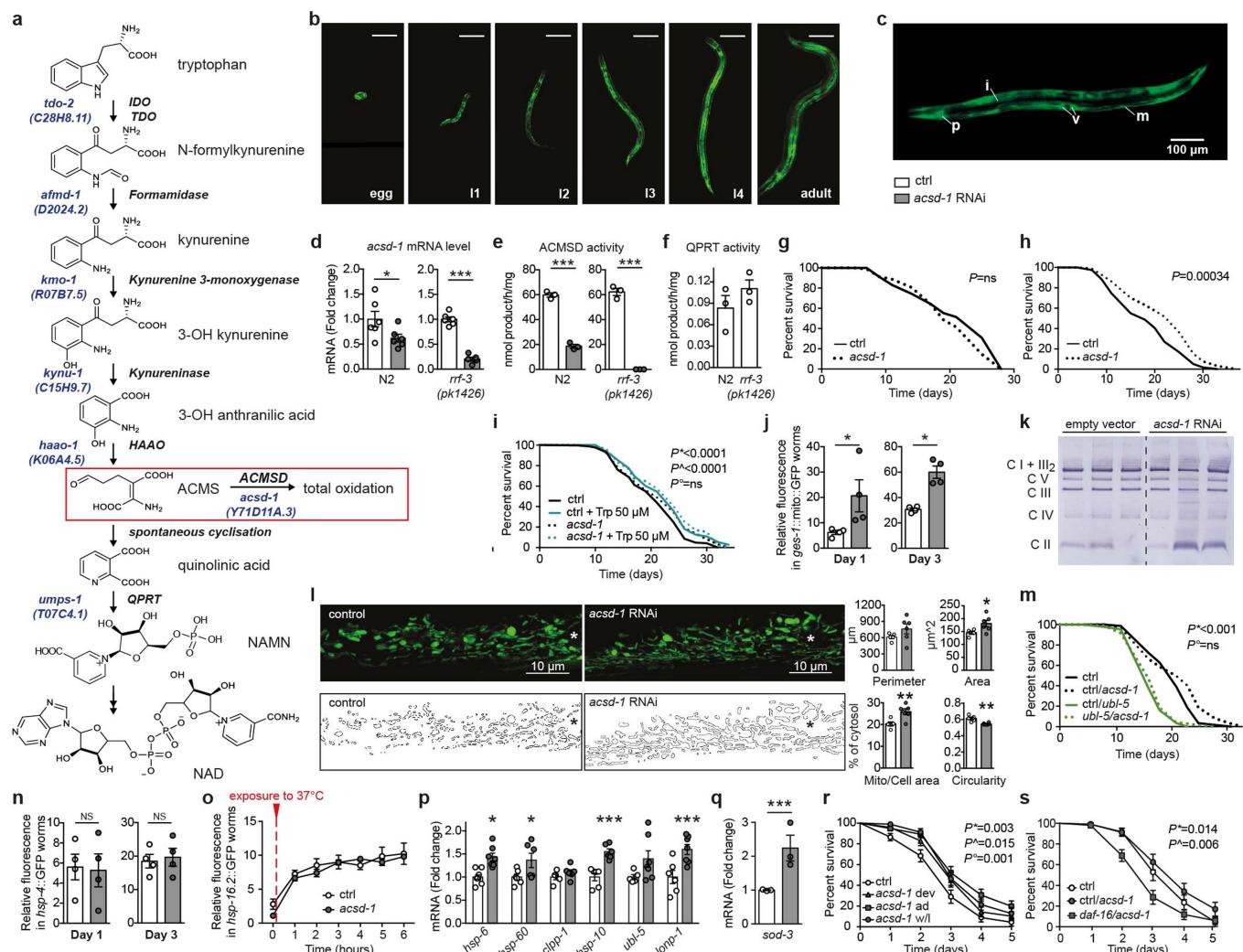
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The authors declare that all the data supporting the findings of this study are available from the corresponding author upon request.

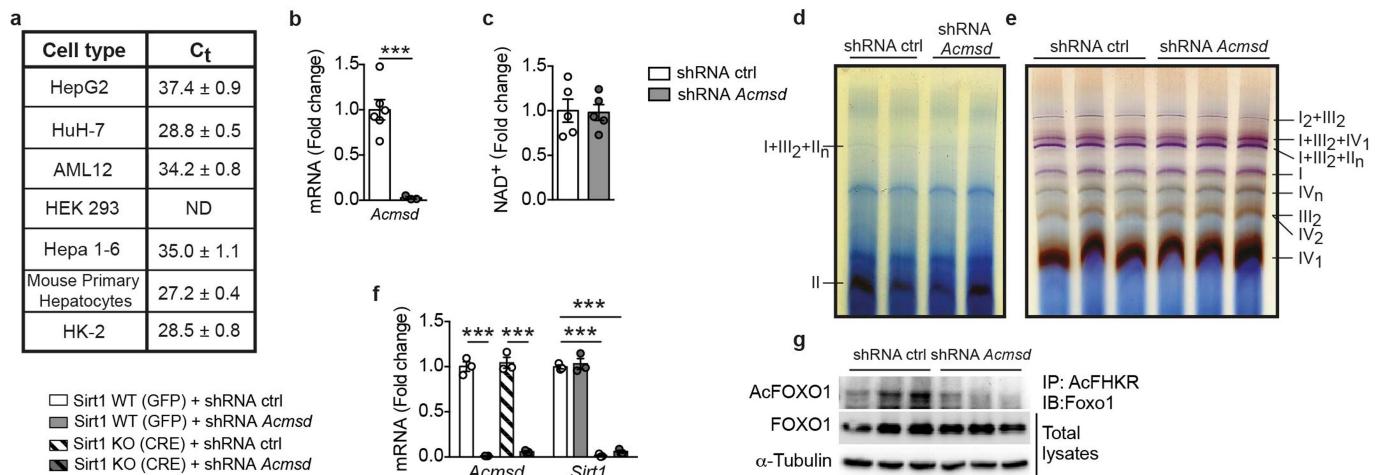
34. Kamath, R. S., Martinez-Campos, M., Zipperlen, P., Fraser, A. G. & Ahringer, J. Effectiveness of specific RNA-mediated interference through ingested double-stranded RNA in *Caenorhabditis elegans*. *Genome Biol.* **12**, RESEARCH0002 (2001).
35. Mouchiroud, L. et al. Pyruvate imbalance mediates metabolic reprogramming and mimics lifespan extension by dietary restriction in *Caenorhabditis elegans*. *Aging Cell* **10**, 39–54 (2011).
36. Mouchiroud, L. et al. The Movement Tracker: a flexible system for automated movement analysis in invertebrate model organisms. *Curr. Protoc. Neurosci.* **77**, 8.37.1–8.37.21 (2016).
37. Zamporlini, F. et al. Novel assay for simultaneous measurement of pyridine mononucleotides synthesizing activities allows dissection of the NAD⁺ biosynthetic machinery in mammalian cells. *FEBS J.* **281**, 5104–5119 (2014).
38. Yang, T. & Sauve, A. A. NAD metabolism and sirtuins: metabolic regulation of protein deacetylation in stress and toxicity. *AAPS J.* **8**, E632–E643 (2006).
39. Oosterveer, M. H. et al. LRH-1-dependent glucose sensing determines intermediary metabolism in liver. *J. Clin. Invest.* **122**, 2817–2826 (2012).

40. Jha, P., Wang, X. & Auwerx, J. Analysis of mitochondrial respiratory chain supercomplexes using blue native polyacrylamide gel electrophoresis (BN-PAGE). *Curr. Protoc. Mouse Biol.* **6**, 1–14 (2016).
41. Lagouge, M. et al. Resveratrol improves mitochondrial function and protects against metabolic disease by activating SIRT1 and PGC-1 α . *Cell* **127**, 1109–1122 (2006).
42. Jha, P. et al. Role of adipose tissue in methionine-choline-deficient model of non-alcoholic steatohepatitis (NASH). *Biochim. Biophys. Acta* **1842**, 959–970 (2014).
43. Schock-Kusch, D. et al. Transcutaneous measurement of glomerular filtration rate using FITC-sinistrin in rats. *Nephrol. Dial. Transplant.* **24**, 2997–3001 (2009).
44. Schreiber, A. et al. Transcutaneous measurement of renal function in conscious mice. *Am. J. Physiol. Renal Physiol.* **303**, F783–F788 (2012).
45. Melnikov, V. Y. et al. Neutrophil-independent mechanisms of caspase-1- and IL-18-mediated ischemic acute tubular necrosis in mice. *J. Clin. Invest.* **110**, 1083–1091 (2002).



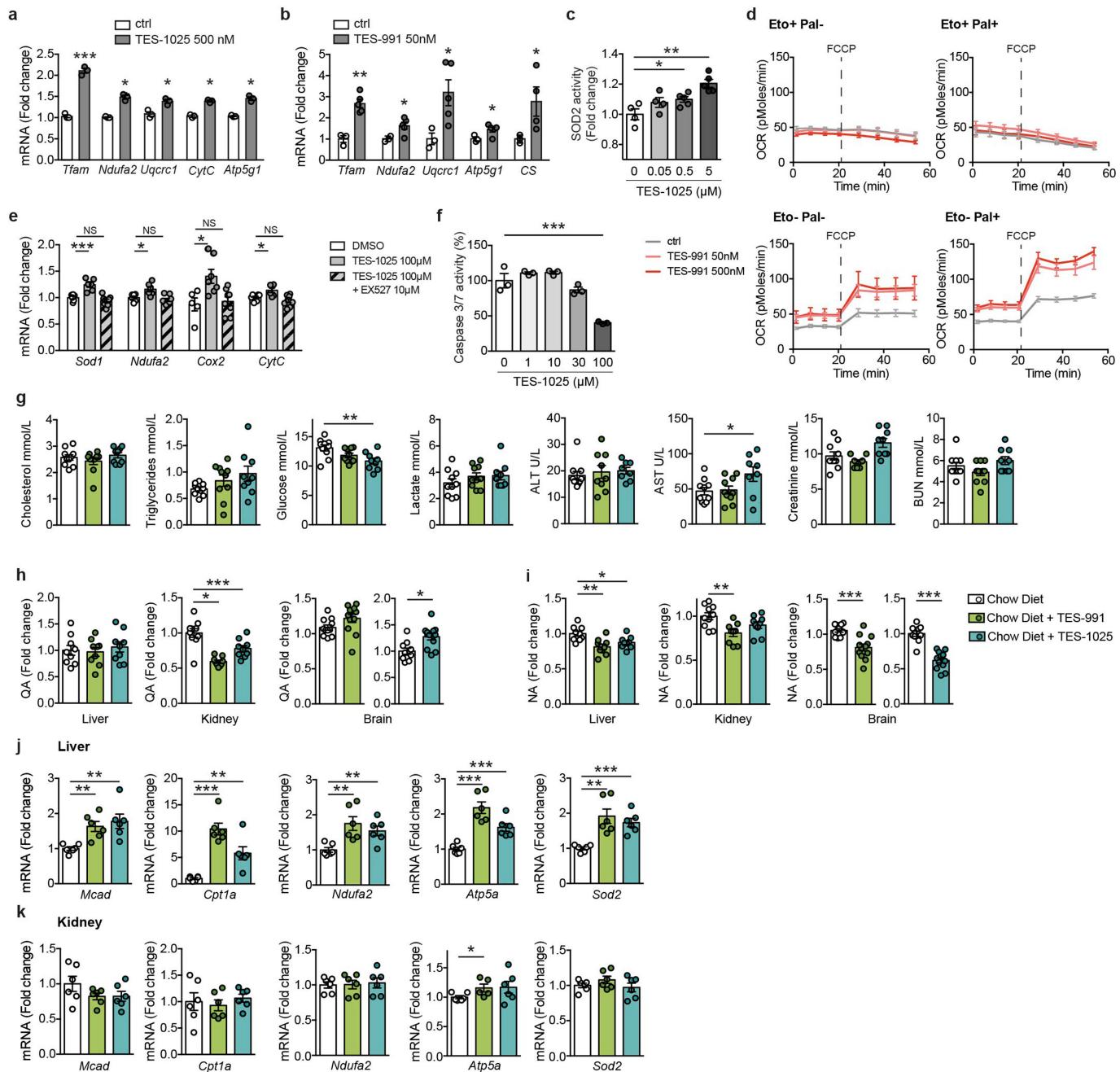
Extended Data Fig. 1 | acsd-1 LOF improves NAD⁺ levels, mitochondrial function, and lifespan through de novo synthesis in *C. elegans*. **a**, De novo synthesis of NAD⁺ from tryptophan. Names of the worm's orthologues are in blue. **b**, *acsd-1* expression pattern across different developmental stages in wild-type worms expressing extrachromosomal array of *acsd-1*::GFP transgene. Scale bar, 100 μ m. **c**, *acsd-1* expression pattern in adult wild-type worms expressing extrachromosomal array of *acsd-1*::GFP transgene. i, intestine; m, muscle; p, pharynx; v, vulva. **d**, *acsd-1* mRNA levels in wild-type and *rrf-3*(*pk1426*) mutants ($n = 6$, each n represents a pool of ~ 600 worms). **e**, ACSD-1 activity in control (empty vector) versus *acsd-1* RNAi-fed worms quantified in both wild-type and *rrf-3* mutants ($n = 3$, where each n represents a pool of $\sim 3,600$ worms) with compensation for negative controls. **f**, QPRT-like activity can be detected in both wild-type worms and *rrf-3* mutants ($n = 3$, each n represents a pool of $\sim 3,600$ worms). **g**, **h**, Effects of *acsd-1* knockdown throughout the entire life on N2 (**g**) and *rrf-3* mutant (**h**) worm lifespan. **i**, Lifespan of *rrf-3*(*pk1426*) mutants exposed to control or *acsd-1* RNAi upon tryptophan supplementation. P^* , ctrl versus ctrl + Trp 50 μ M; P^{\wedge} , ctrl versus *acsd-1* RNAi; P^o , ctrl + Trp 50 μ M versus *acsd-1* RNAi + Trp 50 μ M. **j**, Quantification of GFP signal in *ges-1*::mito::GFP reporter strain, expressing mitochondria-targeted GFP in the intestine at day 1 and 3 of adulthood ($n = 4$, each n represents a pool of 20 worms). **k**, Blue native PAGE on mitochondria extracted from *rrf-3* mutant worms fed with either empty vector or *acsd-1* RNAi bacteria at day 2 of adulthood ($n = 3$, each n represents mitochondria extracted from a pool of $\sim 10,000$ worms). **l**, Mitochondrial morphology in the *myo-3*::mito::GFP reporter strain fed with control or *acsd-1* RNAi. Stars

represent nuclei. Scoring includes the total perimeter of the mitochondrial network, its total area, the area occupied by the mitochondria within the cell and the circularity assessment, in which 1 is a perfect circle and 0 is a line ($n = 6$ worms). **m**, Epistasis between RNAi for *acsd-1* and the UPR^{mt} regulator, *ubl-5*. P^* , ctrl versus ctrl/*acsd-1* RNAi; P^o , ctrl/*ubl-5* RNAi versus *ubl-5*/*acsd-1* RNAi. **n**, Quantification of the GFP signal in *hsp-4*::GFP reporter strain ($n = 4$, each n represents a pool of 20 worms) at day 1 and 3 of adulthood. **o**, Quantification of the GFP signal in *hsp-16.2*::GFP reporter strain ($n = 4$, each n represents a pool of 20 worms). After the first time point sampled at 20 °C, worms were exposed to 37 °C, and the measurement was repeated every hour for 6 h. **p**, Expression of UPR^{mt} genes in worms at day 2 of adulthood fed with control or *acsd-1* RNAi ($n = 6$, each n represents a pool of ~ 600 worms). **q**, Expression of *sod-3* mRNA at day 1 of adulthood in control or *acsd-1* RNAi-fed worms ($n = 3$, each n represents a pool of ~ 600 worms). **r**, Survival of wild-type (N2) worms exposed to 4 mM paraquat starting at the L4 stage, in which the knockdown of *acsd-1* was performed at different life stages. P^* , ctrl versus *acsd-1* RNAi whole life; P^{\wedge} , ctrl versus *acsd-1* RNAi development; P^o , ctrl versus *acsd-1* RNAi adulthood. **s**, Epistasis between RNAi for *acsd-1* and *daf-16* in wild-type (N2) worms exposed to 4 mM paraquat. P^* , ctrl versus ctrl/*acsd-1* RNAi; P^{\wedge} , ctrl/*daf-16* RNAi versus *daf-16*/*acsd-1* RNAi. All worm assays, except for *hsp-16.2*::GFP reporter strain, were performed at 20 °C and repeated at least once. Data are mean \pm s.e.m. $*P \leq 0.05$, $**P \leq 0.01$, $***P \leq 0.001$. P values calculated using two-tailed *t*-test (**d**, **e**, **j**, **l**, **n-q**) or log-rank test (**g-i**, **m**, **r**, **s**). For individual P values, see Source Data. For lifespan values, see Extended Data Table 1.



Extended Data Fig. 2 | Pathways activated by *Acmsd* knockdown in worms are conserved in mammalian cells. **a**, *Acmsd* transcript levels reflected by the C_t values in different hepatic and renal cells and cell lines ($n = 4$). C_t values larger than 35 reflect very low transcript levels. **b**, Efficiency of *Acmsd* shRNA in mouse primary hepatocytes 48 h post adenoviral transduction ($n = 6$). **c**, NAD⁺ levels in mitochondria of AML-12 cells transduced with either shRNA control or shRNA against *Acmsd* ($n = 5$). **d**, **e**, Blue native PAGE followed by in-gel activity assay for complex II (blue) (**d**), and complex I (purple) and IV (brown) (**e**) on mitochondria extracted from mouse primary hepatocytes transduced with either shRNA control or shRNA against *Acmsd* for 48 h. The experiment was performed once. **f**, Primary hepatocytes extracted from a *Sirt1*^{L2/L2}

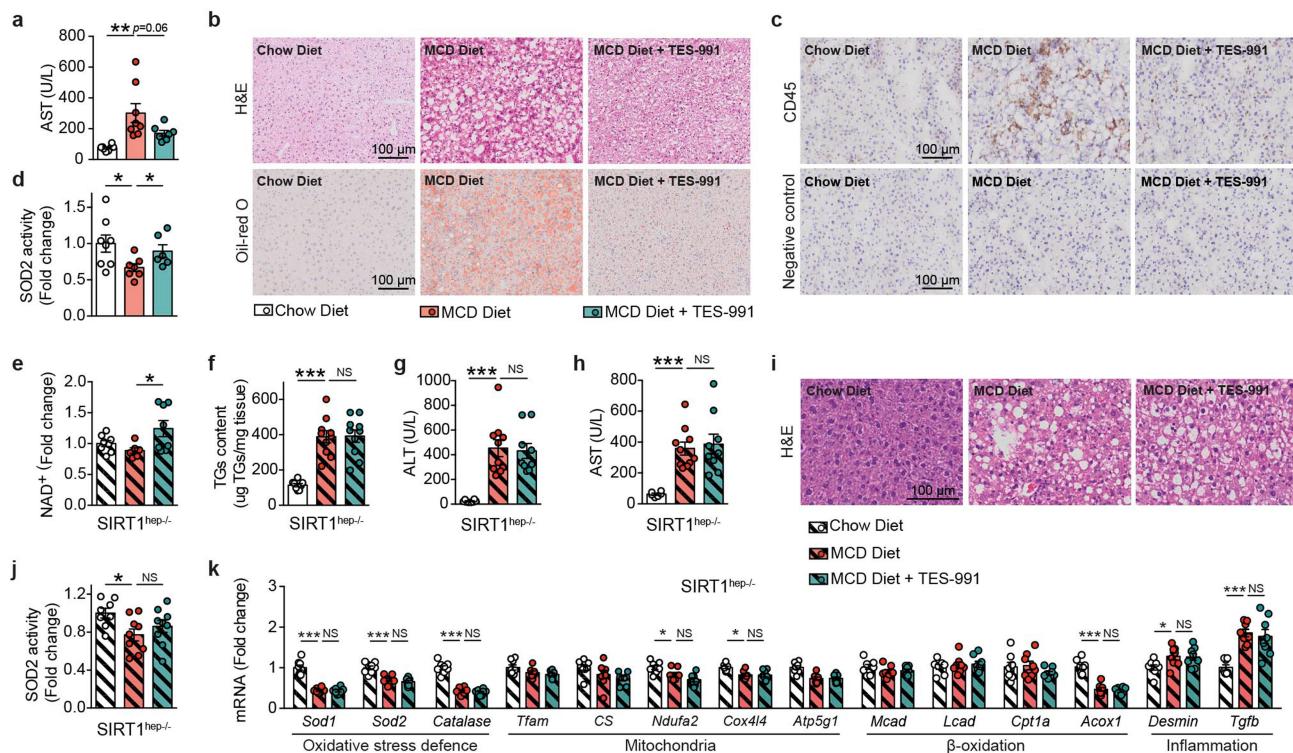
mouse were transduced either with an adenovirus encoding GFP (wild-type condition) or the Cre recombinase to generate *Sirt1* knockout primary hepatocytes. These hepatocytes were exposed to an shRNA targeting a random sequence or shRNA targeting *Acmsd*. Transcript levels of *Acmsd* and *Sirt1* ($n = 3$). **g**, FOXO1 acetylation levels in mouse primary hepatocytes transduced with either shRNA control or shRNA against *Acmsd* for 48 h. The experiment was independently performed twice. Data are mean \pm s.e.m.; each n represents a biologically independent sample. $*P \leq 0.05$, $**P \leq 0.01$, $***P \leq 0.001$. P values calculated using two-tailed *t*-test. For gel source images see Supplementary Fig. 1. For individual P values, see Source Data.



Extended Data Fig. 3 | Pharmacological inhibition of ACMSD has similar effects to genetic downregulation. **a, b**, mRNA levels of mitochondrial genes in mouse primary hepatocytes treated for 24 h with DMSO or TES-1025 (**a**) or TES-991 (**b**), at the indicated concentrations ($n = 3$). **c**, SOD2 activity in mouse primary hepatocytes treated for 24 h with DMSO or TES-1025, at indicated concentrations ($n = 4$). **d**, Fatty acid oxidation assessed in mouse primary hepatocytes treated with DMSO or TES-991 for 24 h at the indicated concentrations ($n = 5$). FCCP (2 μ M) was used as an uncoupler to reach maximal respiration. **e**, mRNA levels of mitochondrial genes in HK-2 cells after 24 h of treatment with TES-1025 or TES-991 in combination with SIRT1 inhibitor, EX527, at the indicated concentrations ($n = 5–8$). **f**, Apoptosis rate in HK-2 cells assessed 16 h after addition of 50 μ M cisplatin by caspase-3/7 activity. TES-1025 was

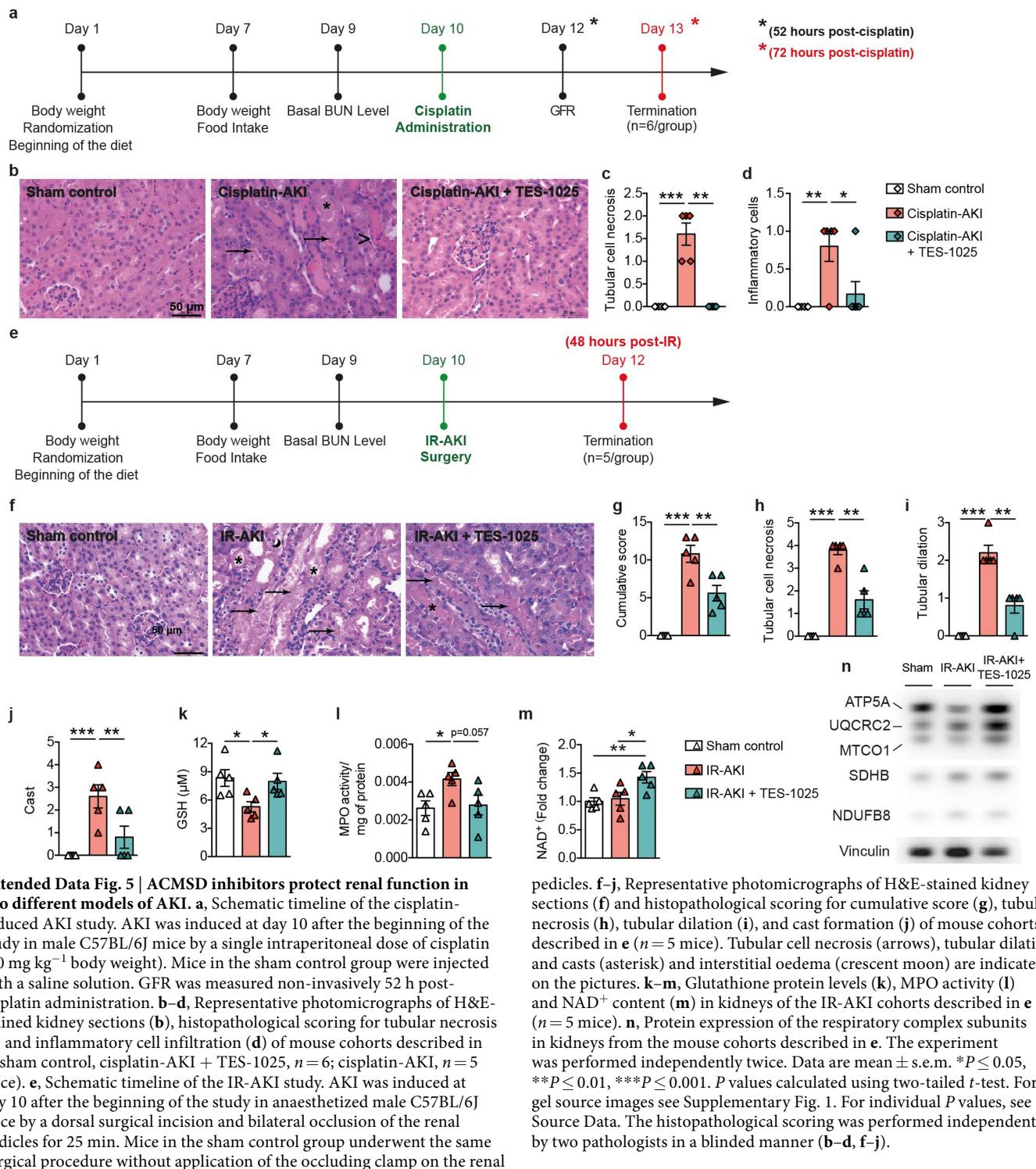
added simultaneously with the cisplatin. **g**, Biochemical analysis of plasma from mice fed with chow diet or chow diet supplemented with TES-991 or TES-1025 at 15 mg kg $^{-1}$ body weight day $^{-1}$ (ctrl and TES-991, $n = 10$; TES-1025, $n = 9$ mice). **h, i**, Quinolinic acid (QA) (**h**) and nicotinic acid (NA) (**i**) levels in livers ($n = 9$), kidneys ($n = 10$; TES-991, TES-1025, $n = 9$ mice) and brains ($n = 11$; TES-991, TES-1025, $n = 12$ mice), from mice fed with control chow diet or chow diet supplemented with TES-991 or TES-1025 at the dose of 15 mg kg $^{-1}$ body weight day $^{-1}$.

j, k, mRNA levels of β -oxidation, mitochondrial and oxidative stress defence genes in livers (**j**) and kidneys (**k**) of mice described in **g** ($n = 6$ mice). Data are mean \pm s.e.m. * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$. P values calculated using two-tailed t -test (**a–d, g–k**) or one-way ANOVA (**e, f**). For individual P values, see Source Data.



Extended Data Fig. 4 | ACMSD inhibitors protect hepatic function from NAFLD induced by MCD diet. **a**, Plasma aspartate transaminase (AST) levels in 16-week-old C57BL/6J male mice fed for 2.5 weeks with control diet, MCD diet or MCD diet supplemented with 15 mg kg⁻¹ day⁻¹ TES-991 ($n = 8$ mice). **b**, Representative photomicrographs of liver tissues stained with H&E or Oil red O from the mouse cohorts described in **a**. The experiment was performed twice independently. **c**, Representative photomicrographs of liver tissues from the mouse cohorts described in **a**, stained with CD45 and the corresponding negative control. The experiment was performed twice independently. **d**, Hepatic SOD2 activity in mouse cohorts described in **a** (chow diet, $n = 8$; MCD diet, $n = 7$; MCD diet + TES-991, $n = 6$ mice). **e–h**, Liver NAD⁺ (**e**), triglyceride content (**f**), plasma ALT (**g**) and AST (**h**) levels in congenic C57BL/6J *Sirt1*^{hep-/-} mice that match the mouse cohorts described in **a** regarding age, gender and treatment duration (chow diet, $n = 8$; MCD diet, MCD diet + TES-991, $n = 10$ mice). **i**, Representative photomicrographs of liver tissues stained with H&E from the *Sirt1*^{hep-/-} mice described in **e–h**. The experiment was performed once. **j**, Hepatic SOD2 activity in congenic C57BL/6J *Sirt1*^{hep-/-} mice described in **e–h** (chow diet, $n = 8$; MCD diet, MCD diet + TES-991, $n = 9$ mice). **k**, mRNA levels of oxidative stress defence, mitochondrial, β -oxidation, inflammatory and fibrosis genes in livers of *Sirt1*^{hep-/-} mice ($n = 8$ mice). Data are mean \pm s.e.m. * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$. P values calculated using two-tailed *t*-test. For individual P values, see Source Data.

plasma ALT (**g**) and AST (**h**) levels in congenic C57BL/6J *Sirt1*^{hep-/-} mice that match the mouse cohorts described in **a** regarding age, gender and treatment duration (chow diet, $n = 8$; MCD diet, MCD diet + TES-991, $n = 10$ mice). **i**, Representative photomicrographs of liver tissues stained with H&E from the *Sirt1*^{hep-/-} mice described in **e–h**. The experiment was performed once. **j**, Hepatic SOD2 activity in congenic C57BL/6J *Sirt1*^{hep-/-} mice described in **e–h** (chow diet, $n = 8$; MCD diet, MCD diet + TES-991, $n = 9$ mice). **k**, mRNA levels of oxidative stress defence, mitochondrial, β -oxidation, inflammatory and fibrosis genes in livers of *Sirt1*^{hep-/-} mice ($n = 8$ mice). Data are mean \pm s.e.m. * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$. P values calculated using two-tailed *t*-test. For individual P values, see Source Data.



Extended Data Fig. 5 | ACMSD inhibitors protect renal function in two different models of AKI. **a**, Schematic timeline of the cisplatin-induced AKI study. AKI was induced at day 10 after the beginning of the study in male C57BL/6J mice by a single intraperitoneal dose of cisplatin (20 mg kg⁻¹ body weight). Mice in the sham control group were injected with a saline solution. GFR was measured non-invasively 52 h post-cisplatin administration. **b-d**, Representative photomicrographs of H&E-stained kidney sections (**b**), histopathological scoring for tubular necrosis (**c**) and inflammatory cell infiltration (**d**) of mouse cohorts described in **a** (sham control, cisplatin-AKI + TES-1025, $n=6$; cisplatin-AKI, $n=5$ mice). **e**, Schematic timeline of the IR-AKI study. AKI was induced at day 10 after the beginning of the study in anaesthetized male C57BL/6J mice by a dorsal surgical incision and bilateral occlusion of the renal pedicles for 25 min. Mice in the sham control group underwent the same surgical procedure without application of the occluding clamp on the renal

pedicles. **f-j**, Representative photomicrographs of H&E-stained kidney sections (**f**) and histopathological scoring for cumulative score (**g**), tubular necrosis (**h**), tubular dilation (**i**), and cast formation (**j**) of mouse cohorts described in **e** ($n=5$ mice). Tubular cell necrosis (arrows), tubular dilation and casts (asterisk) and interstitial oedema (crescent moon) are indicated on the pictures. **k-m**, Glutathione protein levels (**k**), MPO activity (**l**) and NAD⁺ content (**m**) in kidneys of the IR-AKI cohorts described in **e** ($n=5$ mice). **n**, Protein expression of the respiratory complex subunits in kidneys from the mouse cohorts described in **e**. The experiment was performed independently twice. Data are mean \pm s.e.m. * $P\leq 0.05$, ** $P\leq 0.01$, *** $P\leq 0.001$. P values calculated using two-tailed t -test. For gel source images see Supplementary Fig. 1. For individual P values, see Source Data. The histopathological scoring was performed independently by two pathologists in a blinded manner (**b-d**, **f-j**).

Extended Data Table 1 | Summary of *C. elegans* lifespan experiments

Conditions		Cumulative statistics				Statistics of individual expts				
Strain	Treatment/rnai	No. of expts.	Mean lifespan [days] (treatment/ctrl)	Variation [%] compared to control	P-value	No. of animals (treatment/ctrl)	Mean lifespan [days] (treatment/ctrl)	Variation [%] compared to control	P-value	No. of animals (treatment/ctrl)
N2	acsd-1	2	20.5/20.5 (±0.47/0.41)	0	0.3	116/124	20.2/20.9	-3.35	0.303	43/44
							20.6/20.2	+1.98	0.582	73/80
<i>rrf-3</i> (pk1426)	acsd-1	2	21.4/18.5 (±0.56/0.56)	+15.7	0.00034	137/118	22.0/18.6	+18.3	0.004	70/66
							20.9/18.4	+13.6	0.042	67/52
<i>rrf-3</i> (pk1426)	acsd-1 1/2	4	20.2/18.5 (±0.35/0.32)	+9.2	<0.0001	262/246	18.6/17.0	+9.4	0.047	61/57
							19.6/17.6	+11.4	0.03	59/64
							20.2/18.2	+11.0	0.009	63/61
							21.4/20.6	+3.9	0.036	79/64
<i>rrf-3</i> (pk1426)	sir-2.1 1/2	1	17.4/18.1 (±0.51/0.62)	-3.9	0.32	61/61	17.4/18.1	-3.9	0.32	61/61
<i>rrf-3</i> (pk1426)	sir-2.1/acsd-1	1	17.6/18.1 (±0.56/0.62)	-2.8	0.67	66/61	17.6/18.1	-2.8	0.67	66/61
			17.6/17.4 ^a (±0.56/0.51) ^a	+1.1 ^a	0.57 ^a	66/61 ^a				
<i>rrf-3</i> (pk1426)	atfs-1 1/2	2	17.1/17.9 (±0.45/0.46)	-4.5	0.42	135/125	16.8/17.6	-4.5	0.42	69/64
							17.4/18.2	-4.4	0.67	66/61
<i>rrf-3</i> (pk1426)	atfs-1/acsd-1	2	18.0/17.9 (±0.48/0.46)	+0.6	0.72	117/125	18.2/17.6	+3.4	0.58	55/64
			18.0/17.1 ^b (±0.48/0.45) ^b	+5.2 ^b	0.25 ^b	117/135 ^b	17.9/18.2	-1.6	0.94	62/61
<i>rrf-3</i> (pk1426)	ubl-5 1/2	1	17.0/20.6 (±0.3/0.45)	-17.5	<0.0001	68/64	17.0/20.6	-17.5	<0.0001	68/64
<i>rrf-3</i> (pk1426)	ubl-5/acsd-1	1	17.2/20.6 (±0.34/0.45)	-16.5	<0.0001	74/64	17.2/20.6	-16.5	<0.0001	74/64
			17.2/17.0 ^c (±0.34/0.3) ^c	+1.2 ^c	0.36 ^c	74/68 ^c				
<i>rrf-3</i> (pk1426)	daf-16 1/2	2	15.0/17.8 (±0.41/0.52)	-15.7	<0.0001	80/126	13.4/17.0	-21.2	<0.0001	39/60
							16.4/18.6	-11.8	0.01	41/66
<i>rrf-3</i> (pk1426)	daf-16/acsd-1	2	14.1/17.8 (±0.34/0.52)	-20.8	<0.0001	99/126	13.0/17.0	-23.5	<0.0001	49/60
			14.1/15.0 ^d (±0.34/0.4) ^d	-6.0 ^d	0.06 ^d	99/80 ^d	15.1/18.6	-18.8	<0.0001	50/66
Paraquat experiments										
<i>rrf-3</i> (pk1426)	acsd-1 (whole life)	3	7.45/6.85 (±0.14/0.17)	+8.0	0.0034	110/137	8.97/8.57	+4.7	0.014	42/64
							6.42/5.71	+12.4	0.028	47/48
							6.11/5.33	+14.7	0.025	24/35
<i>rrf-3</i> (pk1426)	acsd-1 1/2	2	6.19/5.84 (±0.13/0.15)	+6.48	0.0142	38/44	3.9/3.73	+4.56	0.334	30/32
							6.97/6.68	+4.34	0.043	8/12
<i>rrf-3</i> (pk1426)	daf-16/acsd-1	1	3.07/3.73 (±0.14/0.15)	-21.49	0.006	47/32	3.07/3.73	-21.49	0.006	47/32
<i>rrf-3</i> (pk1426)	acsd-1 (develop)	2	3.55/3.28 (±0.13/0.15)	+8.2	0.0149	73/83	3.43/3.08	+11.4	0.0175	53/52
							3.68/3.48	+5.74	0.0279	20/31
<i>rrf-3</i> (pk1426)	acsd-1 (adult)	2	3.68/3.28 (±0.13/0.15)	+12.2	0.0013	69/83	3.7/3.1	+21.4	0.00082	51/52
							3.66/3.48	+5.17	0.005	18/31
Tryptophan supplementation experiments										
<i>rrf-3</i> (pk1426)	ev/Trp 50μM	2	22.3/20.1 (±0.51/0.40)	+10.9	<0.0001	129/173	22.4/20.6	+8.73	0.025	71/91
							22.2/20.4	+8.82	0.001	58/78
<i>rrf-3</i> (pk1426)	acsd-1/Trp 50μM	2	22.7/20.1 (±0.50/0.40)	+12.9	<0.0001	143/173	23.2/20.6	+12.6	0.001	62/91
			22.7/22.3 ^e (±0.50/0.51) ^e	+1.79 ^e	0.81 ^e	129/173 ^e	22.2/20.4	+8.82	0.06	81/78

^aVersus *rrf-3* (pk1426) *sir-2.1* 1/2 (1/2 denotes condition in which RNAi was diluted by one half with empty vector).^bVersus *rrf-3* (pk1426) *atfs-1* 1/2.^cVersus *rrf-3* (pk1426) *ubl-5* 1/2.^dVersus *rrf-3* (pk1426) *daf-16* 1/2.^eVersus *rrf-3* (pk1426) *ev*/Trp 50 μM.

Errors are represented as s.e.m. Survival analyses were performed using the Kaplan–Meier method and the significance of differences between survival curves was calculated using the log-rank test (two-sided).

Antibody and TLR7 agonist delay viral rebound in SHIV-infected monkeys

Erica N. Borducchi^{1,6}, Jinyan Liu^{1,6}, Joseph P. Nkolola^{1,6}, Anthony M. Cadena^{1,6}, Wen-Han Yu², Stephanie Fischinger², Thomas Broge², Peter Abbink¹, Noe B. Mercado¹, Abishek Chandrashekhar¹, David Jetton¹, Lauren Peter¹, Katherine McMahan¹, Edward T. Moseley¹, Elena Bekerman³, Joseph Hesselgesser³, Wenjun Li⁴, Mark G. Lewis⁵, Galit Alter², Romas Gelezunas³ & Dan H. Barouch^{1,2*}

The latent viral reservoir is the critical barrier for the development of a cure for HIV-1 infection. Previous studies have shown direct antiviral activity of potent HIV-1 Env-specific broadly neutralizing antibodies (bNAbs) administered when antiretroviral therapy (ART) was discontinued, but it remains unclear whether bNAbs can target the viral reservoir during ART. Here we show that administration of the V3 glycan-dependent bNAb PGT121 together with the Toll-like receptor 7 (TLR7) agonist vesatolimod (GS-9620) during ART delayed viral rebound following discontinuation of ART in simian-human immunodeficiency virus (SHIV)-SF162P3-infected rhesus monkeys in which ART was initiated during early acute infection. Moreover, in the subset of monkeys that were treated with both PGT121 and GS-9620 and that did not show viral rebound after discontinuation of ART, adoptive transfer studies and CD8-depletion studies also did not reveal virus. These data demonstrate the potential of bNAb administration together with innate immune stimulation as a possible strategy for targeting the viral reservoir.

The viral reservoir in latently infected CD4⁺ T lymphocytes^{1–4} is responsible for viral rebound in the vast majority of HIV-1-infected individuals who stop taking ART and represents the key challenge to a cure for HIV-1 infection^{5,6}. Multiple strategies for an HIV-1 cure are currently being pursued. One hypothesis is that activation of reservoir cells may render them more susceptible to immune-mediated destruction^{7–9}, but, to our knowledge, direct evidence that this strategy can reduce the viral reservoir in vivo has not previously been reported.

Potent HIV-1-specific bNAbs have been shown to reduce viraemia in untreated, chronically SHIV-infected rhesus monkeys^{10–12} and in HIV-1-infected humans^{13,14}. Moreover, bNAbs have been reported to delay viral rebound in HIV-1-infected humans when the antibodies were administered at the time of discontinuation of ART^{15,16}. These studies demonstrate that bNAbs can exert direct antiviral activity, but whether they can target the viral reservoir during ART suppression remains to be determined. Such an experiment would require that bNAbs no longer be present at therapeutic levels when ART is discontinued. To explore this concept, we evaluated the capacity of the potent neutralizing antibody PGT121^{17,18} and the TLR7 agonist vesatolimod (GS-9620)^{19,20} to target the viral reservoir in ART-suppressed, SHIV-SF162P3-infected rhesus monkeys.

Study design

We infected 44 Indian-origin rhesus monkeys (*Macaca mulatta*) by the intrarectal route with our rhesus peripheral blood mononuclear cell (PBMC)-derived stock of SHIV-SF162P3²¹ and initiated daily subcutaneous administration of a pre-formulated ART cocktail (tenofovir disoproxil fumarate, emtricitabine, dolutegravir)²⁰ on day 7 of infection (Extended Data Fig. 1). Monkeys had median plasma viral loads of 4.2–4.3 log RNA copies per ml (range 2.8–5.8 log RNA copies per ml) on day 7 before initiation of ART (Fig. 1a). Viral loads were balanced among the different groups. Following initiation of ART, plasma viraemia declined to undetectable levels in most monkeys by week 3 and in all monkeys by week 6.

All monkeys were treated with continuous daily ART for 96 weeks and then received one of the following interventions ($n = 11$ monkeys per group; Fig. 1b, Extended Data Fig. 1): sham (group 1), GS-9620 alone (group 2), PGT121 alone (group 3), or both PGT121 and GS-9620 (group 4). In groups 2 and 4, monkeys received ten oral administrations of 0.15 mg kg⁻¹ GS-9620 (Gilead Sciences) every two weeks from weeks 96 to 114. In groups 3 and 4, monkeys received five intravenous infusions of 10 mg kg⁻¹ PGT121 (Catalent Biopharma) every two weeks from weeks 106 to 114. Thus, in the combination group, monkeys first received five doses of GS-9620 alone and then received five doses of both PGT121 and GS-9620. ART was continued for 16 weeks after the final administration of PGT121 and GS-9620 to allow antibody washout before discontinuation of ART at week 130. Viral loads remained undetectable in all monkeys throughout the period of ART suppression with no evidence of viral ‘blips’ to week 130 (Fig. 1b).

Pharmacodynamics and pharmacokinetics

Triggering of TLR7 leads to innate immune activation^{22,23}, and we previously reported that treatment with the TLR7 agonist GS-986 resulted in robust CD4⁺ T cell activation in rhesus monkeys²⁰. In the present study, the related TLR7 agonist GS-9620 similarly led to activation of CD4⁺ T cells, as evidenced by increased CD69 and CD38 expression on CD4⁺ T cells on the day after each GS-9620 administration (Fig. 2a, Extended Data Fig. 2a; $P = 0.001–0.03$, Mann–Whitney tests). Moreover, GS-9620 activated natural killer (NK) cells (Fig. 2b; $P = 0.001–0.01$, Mann–Whitney tests) and monocytes (data not shown), and led to increased plasma levels of IFN α , IL-1RA, I-TAC, eotaxin, MIG, MCP-1, IL-1 β , IL-6 and IP-10 (Extended Data Fig. 2b).

PGT121 was detected in serum after each of the five infusions from weeks 106 to 114 without evidence of induction of anti-drug antibodies (Extended Data Figs. 3, 4a). After the final PGT121 infusion, PGT121 levels fell rapidly, reflecting the short half-life of this human

¹Center for Virology and Vaccine Research, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. ²Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA, USA.

³Gilead Sciences, Foster City, CA, USA. ⁴University of Massachusetts Medical School, Worcester, MA, USA. ⁵Bioqual, Rockville, MD, USA. ⁶These authors contributed equally: Erica N. Borducchi, Jinyan Liu, Joseph P. Nkolola, Anthony M. Cadena. *e-mail: dbarouch@bidmc.harvard.edu

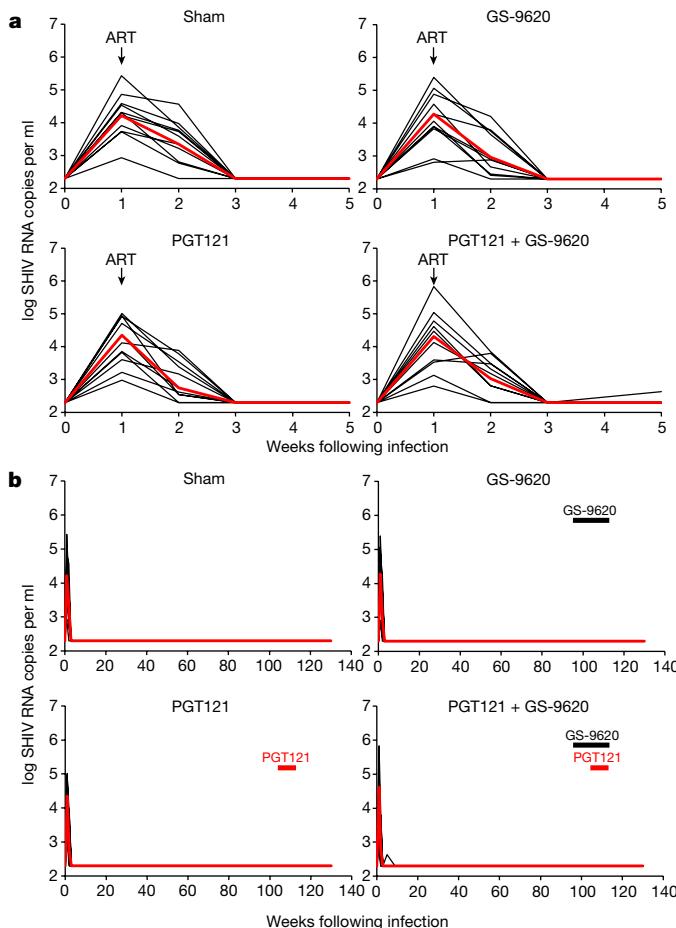


Fig. 1 | Plasma viral loads after infection with SHIV-SF162P3 and before discontinuation of ART. We infected 44 rhesus monkeys with SHIV-SF162P3 at week 0 and initiated ART at week 1 (day 7) ($n=11$ monkeys per group). Administration of GS-9620 and infusions of PGT121 were performed from weeks 96 to 114. ART was discontinued at week 130. Plasma viral loads are shown for weeks 0–5 (a) and weeks 0–130 (b). Data are shown as log SIV RNA copies per ml (limit of detection 2.3 log RNA copies per ml). Red lines indicate median values.

antibody in monkeys¹⁰. Serum PGT121 declined to undetectable levels ($<0.5 \mu\text{g ml}^{-1}$) in most monkeys by week 120 and in all monkeys by week 122 (Extended Data Fig. 4a). PGT121 was also undetectable in cell lysates and supernatants from lymph nodes and colorectal biopsies from week 120 (Extended Data Fig. 4b). We previously defined a PGT121 level of $1 \mu\text{g ml}^{-1}$ as the threshold for viral rebound in SHIV-SF162P3-infected rhesus monkeys for PGT121-mediated virologic suppression¹⁰. Thus, PGT121 levels were below this rebound threshold in peripheral blood and lymphoid and gastrointestinal tissues for 8–10 weeks before discontinuation of ART at week 130.

Viral DNA

Viral DNA was largely undetectable in PBMCs in all groups of monkeys at week 96 and week 120 using quantitative PCR with reverse transcription (RT-PCR) with a sensitivity of 3 DNA copies per 10^6 PBMCs²⁰ (Fig. 3a, b), presumably as a result of the early initiation and the extended duration of suppressive ART. We still detected viral DNA in lymph nodes in the majority of monkeys at week 96 before the interventions, consistent with previous findings in rhesus monkeys treated with ART during acute SIVmac251 infection²⁴. Following the interventions at week 120, we observed similar levels of viral DNA in the sham and GS-9620-alone groups. By contrast, monkeys treated with PGT121 and GS-9620 had lower levels of viral DNA in lymph nodes than did sham controls at week 120 (Fig. 3b; $P=0.004$, Mann–Whitney test). We were unable to detect cellular RNA in all groups, both at week 96 and at week 120

(data not shown). These data suggest that bNAb administration with innate immune stimulation may have reduced the viral reservoir in these monkeys.

Cellular immune responses

Monkeys showed low levels of Gag-, Env-, and Pol-specific cellular immune responses at week 4, and these responses waned by week 96 and week 120 (Extended Data Fig. 5), presumably reflecting a lack of antigen stimulation during the extended period of ART suppression. Multi-parameter intracellular cytokine staining at week 120 corroborated these findings and showed minimal to no detectable CD8⁺ T cell responses in PBMCs in all groups of monkeys (Extended Data Fig. 6a). Lymph node CD8⁺ T cell responses and follicular CXCR5⁺CD8⁺ T cell responses were also marginal in all groups (Extended Data Fig. 6b). In particular, Gag-, Env- and Pol-specific CD8⁺ T cell responses in PBMCs and lymph nodes were not higher in monkeys that received PGT121 than in those that did not, suggesting that there was no ‘vaccinal effect’ following antibody administration in this study.

Viral rebound following ART discontinuation

At week 130, we discontinued ART in all monkeys and monitored plasma viral loads for 196 days to assess viral rebound (Fig. 4a). In the sham control group, 11 of 11 monkeys (100%) rebounded with high peak viral loads and moderate setpoint viral loads typical of SHIV-SF162P3 infection^{20,21}. In the GS-9620-alone group, 10 of 11 monkeys (91%) rebounded, and in the PGT121-alone group, 9 of 11 monkeys (82%) rebounded. In the group treated with both PGT121 and GS-9620, however, only 6 of 11 monkeys (55%) rebounded by day 196 following ART discontinuation. The monkeys treated with PGT121 and GS-9620 that rebounded also showed a 2.6 log reduction in peak viral loads and a 1.5 log reduction in setpoint viral loads as compared with sham controls (Fig. 4b; $P<0.001$, χ^2 test comparing total viral load AUC). All monkeys that rebounded generated robust Gag-, Env- and Pol-specific cellular immune responses following viral rebound, which probably contributed to post-rebound virologic control (Extended Data Fig. 7a).

The median time to viral rebound in the sham group and in the GS-9620-alone group was 21 days, whereas the median time to viral rebound in the group that received PGT121 and GS-9620 was 5.3-fold longer at 112 days (Fig. 5a). Overall, the group that received PGT121 and GS-9620 demonstrated significantly delayed viral rebound compared with the sham group, and the PGT121-alone group showed a detectable but more modest effect (Fig. 5a, b; $P=0.0001$, Kruskall–Wallis test comparing all groups; $P<0.001$, censored Poisson regression model comparing PGT121+GS-9620 versus sham, PGT121 alone versus sham, and PGT121+GS-9620 versus PGT121 alone). These findings demonstrate that treatment with PGT121 and GS-9620 was superior to PGT121 alone in delaying viral rebound following ART discontinuation.

Monkeys treated with PGT121 and GS-9620 that did not rebound by day 196 following ART discontinuation had lower peak pre-ART plasma viral loads at week 1 of infection than those that did rebound (Fig. 5c; $P=0.03$, Mann–Whitney test). Moreover, pre-ART viral loads correlated inversely with time to rebound in this group (Fig. 5c; $P=0.03$, $R=-0.62$, Spearman rank-correlation test). Similar trends were observed in the monkeys that did not rebound in the other groups (Extended Data Fig. 7b). These data suggest that the extent of viral exposure before ART initiation was a key determinant of the therapeutic efficacy of PGT121 and GS-9620, probably by limiting initial seeding of the viral reservoir.

Adoptive transfer and CD8 depletion studies

Current viral DNA assays are insufficiently sensitive to predict viral rebound following ART discontinuation, as shown by the rapid rebound in 100% of sham controls (Fig. 4a), despite the absence of detectable viral DNA in PBMCs and lymph nodes in a subset of these monkeys at week 120 (Fig. 3b). These findings are also consistent with clinical observations that have shown viral rebound even in patients

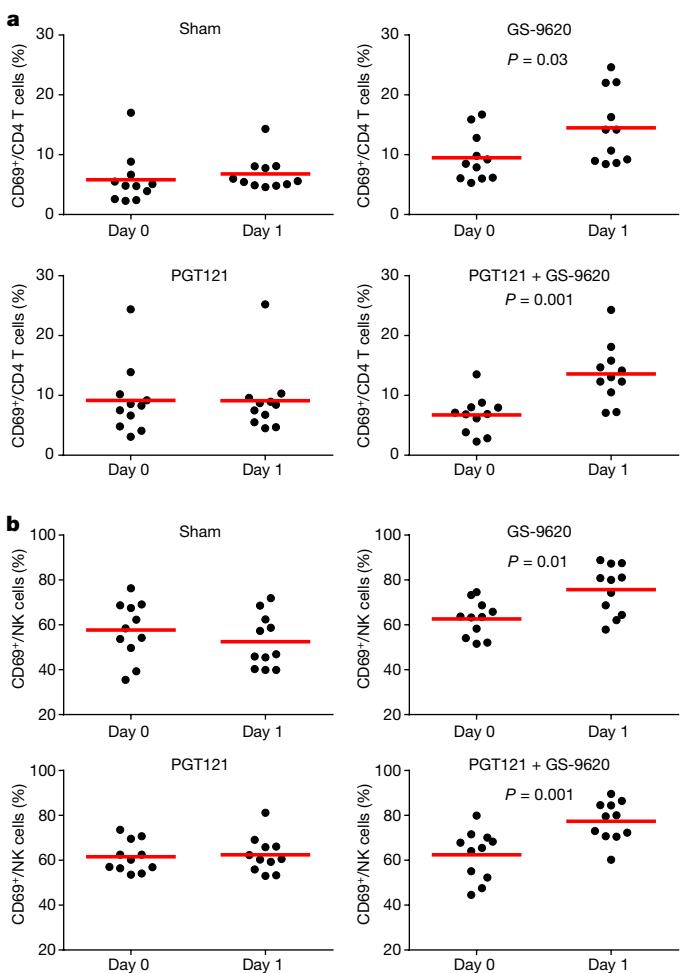


Fig. 2 | Cellular immune activation following administration of GS-9620 and before discontinuation of ART. Activation of CD4⁺ T cells (a) and NK cells (b) was assessed by CD69 expression on days 0 and 1 after administration of GS-9620 ($n = 11$ monkeys per group). Representative data are shown following the fifth dose of GS-9620, which was comparable to the other doses. Red horizontal bars indicate median values. P values calculated using two-sided Mann–Whitney tests.

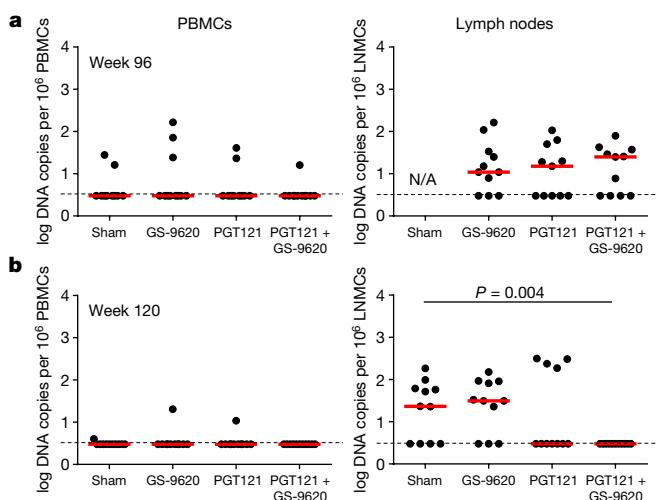


Fig. 3 | Viral DNA before discontinuation of ART. Data are shown as log viral DNA copies per 10^6 cells (limit of detection 3 DNA copies per 10^6 cells) in PBMCs and lymph node mononuclear cells (LNMCs) at week 96 before the interventions (a) and at week 120 following the interventions (b) ($n = 11$ monkeys per group). Red horizontal bars indicate median values. P values calculated using two-sided Mann–Whitney tests.

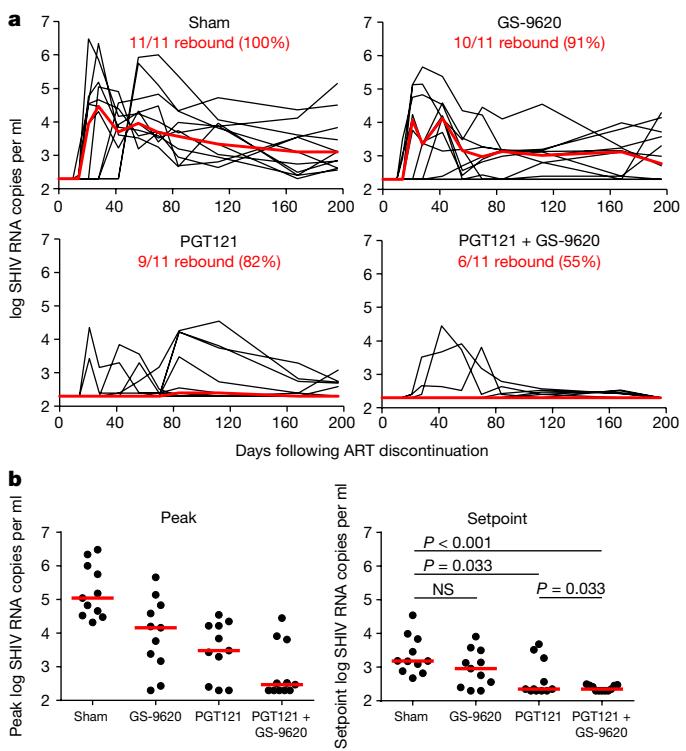


Fig. 4 | Viral loads following ART discontinuation. a, Plasma viral loads are shown for 196 days following ART discontinuation ($n = 11$ monkeys per group). Data are shown as log SIV RNA copies per ml (limit of detection 2.3 log RNA copies per ml). Numbers and percentages of monkeys that show viral rebound by day 196, defined as any confirmed detectable viral load, are shown. b, Summary of peak and setpoint viral loads following ART discontinuation ($n = 11$ monkeys per group). Red lines and horizontal bars indicate median values. P values calculated using χ^2 tests of area under the curve (AUC) total viral RNA following discontinuation of ART. NS, not significant.

with undetectable levels of viral DNA^{25,26}. To assess residual replication-competent virus, we first performed an adoptive transfer study using PBMCs and lymph node mononuclear cells (LNMCs) from five of the seven monkeys that did not rebound and for which we had sufficient LNMCs available from day 140 following ART discontinuation, including four of the five monkeys treated with PGT121 and GS-9620 and one of the two monkeys treated with PGT121 alone. We infused 30 million PBMCs and LNMCs from these five monkeys by intravenous infusion into naïve rhesus monkeys, and we used a similar number of PBMCs and LNMCs from two monkeys treated with PGT121 and GS-9620 that showed transient rebound followed by virologic control as positive controls. All monkeys had undetectable plasma viral RNA on day 140 when PBMCs and LNMCs were collected. Adoptive transfer of cells from the two monkeys that rebounded led to efficient infection of the naïve hosts, as shown by high plasma viral loads of 6.1–6.8 log RNA copies per ml in recipients by day 14–21 following adoptive transfer (Fig. 6a, red lines). These data show that replication-competent virus persisted in these two monkeys despite post-rebound virologic control. By contrast, adoptive transfer of cells from the five monkeys that did not rebound failed to transfer infection to naïve hosts (Fig. 6a, black lines).

We next depleted CD8⁺ T cells and NK cells in the monkeys treated with PGT121 and GS-9620 that had undetectable viral loads on day 196 following ART discontinuation. Monkeys received a single intravenous infusion of 50 mg kg⁻¹ of the anti-CD8 α CDR-grafted rhesus IgG1 antibody MT807R1 (provided by K. Reimann, MassBiologics). Following infusion, all monkeys showed marked CD8 depletion in PBMCs (Extended Data Fig. 8) as well as in lymph nodes and colorectal biopsies (data not shown). In all the monkeys treated with PGT121 and GS-9620 that rebounded following ART discontinuation, plasma

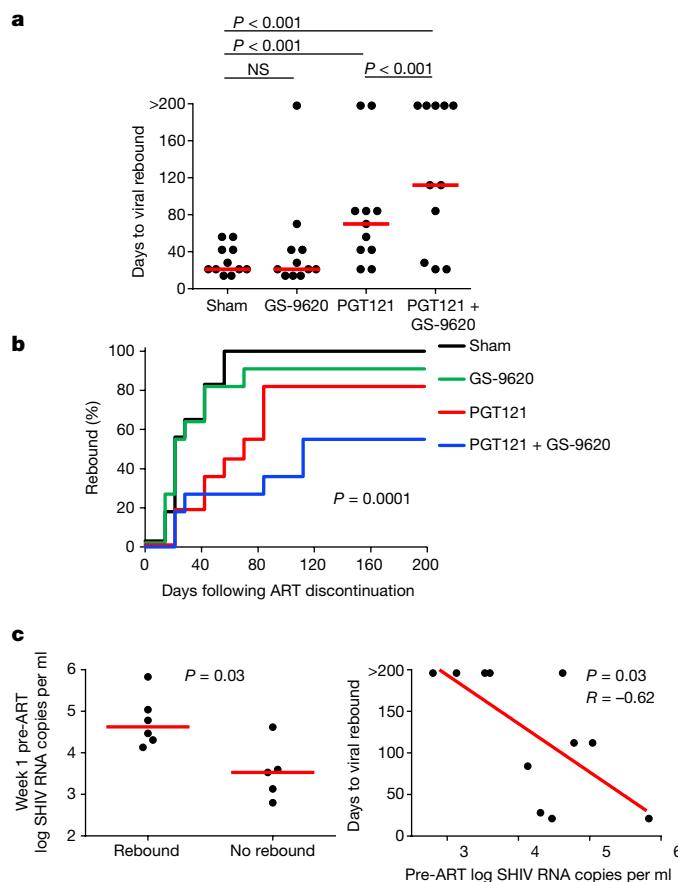


Fig. 5 | Analysis and correlations of viral rebound. **a, b**, Comparison of time to viral rebound among groups as dot plots depicting raw data (**a**) and Kaplan–Meier curves (**b**) ($n=11$ monkeys per group). **c**, Correlation of viral rebound in monkeys treated with PGT121 and GS-9620 with pre-ART week 1 log viral RNA copies per ml ($n=11$). Red horizontal bars indicate median values. P values calculated using a censored Poisson regression model of incidence rate ratios (**a**), a Kruskall–Wallis test (**b**), and a two-sided Spearman rank-correlation test (**c**).

viral loads spiked to 6.0–6.8 log RNA copies per ml by day 7–14 after CD8 depletion (Fig. 6b, red lines). By contrast, in the five monkeys that were treated with PGT121 and GS-9620 and did not rebound following ART discontinuation, viral loads remained undetectable following CD8 depletion (Fig. 6b, black lines).

Mechanistic studies

We developed a computational model that included cellular activation, innate plasma cytokines, and adaptive immune responses to define the correlates of delayed viral rebound in this study. A model that included NK cell activation correlated best with delayed viral rebound (Extended Data Fig. 9a; $P=5.31 \times 10^{-4}$, $R=0.50$, Spearman rank-correlation test). A similar model involving NK cell and monocyte activation correlated with reduced viral loads following ART discontinuation (Extended Data Fig. 9b; $P=3.52 \times 10^{-6}$, $R=0.64$, Spearman rank-correlation test). These data suggest that activated NK cells and monocytes may have a key role in PGT121-mediated elimination of infected CD4⁺ T cells, consistent with the observed NK cell activation following GS-9620 activation (Fig. 2b).

Finally, we assessed the ability of PGT121, with or without GS-9620, to kill HIV-1-infected CD4⁺ T cells in vitro. GS-9620 led to activation of NK cells and T cells in vitro (Extended Data Fig. 10a), consistent with the in vivo data (Fig. 2, Extended Data Fig. 2). Moreover, the combination of PGT121 and GS-9620 led to optimal killing of HIV-1-infected CD4⁺ T cells in vitro (Extended Data Fig. 10b), consistent with previous studies¹⁹. Together, these data suggest a mechanism by

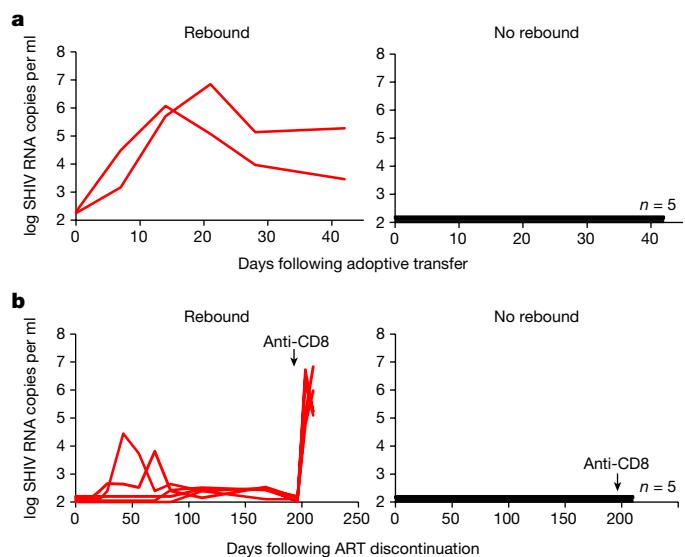


Fig. 6 | Adoptive transfer and CD8 depletion studies. **a**, Plasma viral loads in recipient monkeys following adoptive transfer of 30 million PBMCs and LNMCs from monkeys treated with PGT121 and GS-9620 that exhibited viral rebound and post-rebound virologic control ($n=2$, left, red lines) and monkeys treated with PGT121 and GS-9620 or PGT121 alone that exhibited no viral rebound following discontinuation of ART ($n=5$, which represents 4 from the PGT121 + GS-9620 group and 1 from the PGT121-alone group, right, black lines). **b**, Plasma viral loads in monkeys treated with PGT121 and GS-9620 before and after CD8 depletion (see Methods) in monkeys that exhibited viral rebound ($n=5$, left, red lines) and in monkeys that exhibited no viral rebound following discontinuation of ART ($n=5$, right, black lines). CD8 depletion was performed on day 196 (arrows).

which the TLR7 agonist GS-9620 stimulated innate immunity and activated multiple immune cell subsets in vivo including infected CD4⁺ T cells (Fig. 2, Extended Data Fig. 2), rendering them more susceptible to PGT121-mediated recognition, as well as effector cells such as NK cells and monocytes (Fig. 2b, Extended Data Fig. 9), which may have facilitated PGT121-mediated elimination of these infected CD4⁺ T cells.

Discussion

Our data demonstrate that PGT121 together with the TLR7 agonist GS-9620 delayed viral rebound following discontinuation of ART in acutely treated, SHIV-SF162P3-infected rhesus monkeys. Moreover, five of eleven monkeys that were treated with PGT121 and GS-9620 showed no viral rebound for more than 6 months after cessation of ART and also did not reveal virus by highly sensitive adoptive transfer and CD8 depletion studies. This proof-of-concept study suggests that bNAb administration combined with innate immune stimulation may represent a potential strategy to target the viral reservoir.

Our findings extend previous observations that HIV-1-specific bNAbs have direct antiviral activity in SHIV-infected rhesus monkeys^{10–12} and in HIV-1-infected humans^{13–16}. However, these previous studies did not assess the potential of bNAbs to target the viral reservoir, which would require bNAbs to be administered during ART suppression and no longer be present at therapeutic levels following discontinuation of ART. We previously defined a PGT121 level of 1 μ g ml⁻¹ as the threshold for viral rebound in SHIV-SF162P3-infected rhesus monkeys for PGT121-mediated virologic control¹⁰. In the present study, PGT121 levels were undetectable (below 0.5 μ g ml⁻¹) in peripheral blood, lymph nodes, and colorectal tissue for 8–10 weeks before discontinuation of ART (Extended Data Fig. 4). These findings suggest that the delayed viral rebound in monkeys treated with PGT121 and GS-9620 may reflect reservoir targeting rather than just direct antiviral activity.

The five monkeys in the PGT121- and GS-9620-treated group that showed sustained remission for more than 6 months following discontinuation of ART also did not show evidence of virus in adoptive transfer and CD8 depletion studies (Fig. 6), which are sensitive tests for residual replication-competent virus. Viral rebound can occur in humans following extended periods of ART-free remission^{25,26}. We therefore cannot exclude the possibility that exceedingly low levels of replication-competent virus may still exist in these monkeys. Nevertheless, there was a clear difference between the monkeys that rebounded and those that did not rebound in the adoptive transfer and CD8 depletion studies (Fig. 6).

We hypothesize that the mechanism of the observed effects of PGT121 and GS-9620 involves activation of multiple cell types by GS-9620 followed by efficient binding and elimination of virally infected CD4⁺ T cells by PGT121. GS-9620 activated CD4⁺ T cells and NK cells, both *in vivo* (Fig. 2, Extended Data Fig. 2) and *in vitro* (Extended Data Fig. 10), and activated NK cells and monocytes correlated with delayed viral rebound following ART discontinuation (Extended Data Fig. 9). These data suggest that GS-9620 may have activated latently infected CD4⁺ T cells, possibly rendering them more susceptible to PGT121 binding, and effector cells such as NK cells and monocytes, which may have facilitated antibody-mediated elimination of the infected CD4⁺ T cells. This proposed mechanism predicts that a longer duration of combined PGT121 and GS-9620 therapy might improve therapeutic efficacy. In monkeys treated with PGT121 and GS-9620, we did not detect evidence of a 'vaccinal effect' of increased autologous antigen-specific CD8⁺ T cell responses following bNAb administration (Extended Data Figs. 5, 6), in contrast to a previous study that involved bNAb administration during acute SHIV infection²⁷.

In summary, our data show that bNAb administration combined with innate immune stimulation can delay viral rebound following discontinuation of ART. This study used rhesus monkeys in which ART was initiated on day 7 of acute infection and then received ART continuously for 2.5 years. Moreover, the maximum therapeutic effect was observed in monkeys with the lowest pre-ART viral loads (Fig. 5c). It will therefore probably be far more difficult to achieve similar results in monkeys in which ART is initiated during chronic infection, and thus implications for typical HIV-1-infected humans remain unclear. Nevertheless, our data provide an initial proof-of-concept in primates showing the potential of innate immune activation with immune-based targeting of the viral reservoir.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0600-6>.

Received: 21 April 2018; Accepted: 14 September 2018;

Published online 3 October 2018.

1. Finzi, D. et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**, 1295–1300 (1997).
2. Persaud, D., Zhou, Y., Siliciano, J. M. & Siliciano, R. F. Latency in human immunodeficiency virus type 1 infection: no easy answers. *J. Virol.* **77**, 1659–1665 (2003).
3. Chun, T. W. et al. Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proc. Natl Acad. Sci. USA* **94**, 13193–13197 (1997).
4. Ho, Y. C. et al. Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell* **155**, 540–551 (2013).
5. Finzi, D. et al. Latent infection of CD4⁺ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nat. Med.* **5**, 512–517 (1999).
6. Chun, T. W., Davey, R. T., Jr, Engel, D., Lane, H. C. & Fauci, A. S. Re-emergence of HIV after stopping therapy. *Nature* **401**, 874–875 (1999).

7. Barouch, D. H. & Deeks, S. G. Immunologic strategies for HIV-1 remission and eradication. *Science* **345**, 169–174 (2014).
8. Deeks, S. G. et al. International AIDS Society global scientific strategy: towards an HIV cure 2016. *Nat. Med.* **22**, 839–850 (2016).
9. Shan, L. et al. Stimulation of HIV-1-specific cytolytic T lymphocytes facilitates elimination of latent viral reservoir after virus reactivation. *Immunity* **36**, 491–501 (2012).
10. Barouch, D. H. et al. Therapeutic efficacy of potent neutralizing HIV-1-specific monoclonal antibodies in SHIV-infected rhesus monkeys. *Nature* **503**, 224–228 (2013).
11. Shingai, M. et al. Antibody-mediated immunotherapy of macaques chronically infected with SHIV suppresses viraemia. *Nature* **503**, 277–280 (2013).
12. Julg, B. et al. Virological control by the CD4-binding site antibody N6 in simian-human immunodeficiency virus-infected rhesus monkeys. *J. Virol.* **91**, e00498-17 (2017).
13. Caskey, M. et al. Viraemia suppressed in HIV-1-infected humans by broadly neutralizing antibody 3BNC117. *Nature* **522**, 487–491 (2015).
14. Caskey, M. et al. Antibody 10-1074 suppresses viremia in HIV-1-infected individuals. *Nat. Med.* **23**, 185–191 (2017).
15. Scheid, J. F. et al. HIV-1 antibody 3BNC117 suppresses viral rebound in humans during treatment interruption. *Nature* **535**, 556–560 (2016).
16. Bar, K. J. et al. Effect of HIV antibody VRC01 on viral rebound after treatment interruption. *N. Engl. J. Med.* **375**, 2037–2050 (2016).
17. Walker, L. M. et al. Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature* **477**, 466–470 (2011).
18. Liu, J. et al. Antibody-mediated protection against SHIV challenge includes systemic clearance of distal virus. *Science* **353**, 1045–1049 (2016).
19. Tsai, A. et al. Toll-like receptor 7 agonist GS-9620 induces HIV expression and HIV-specific immunity in cells from HIV-infected individuals on suppressive antiretroviral therapy. *J. Virol.* **91**, e02166-16 (2017).
20. Borducchi, E. N. et al. Ad26/MVA therapeutic vaccination with TLR7 stimulation in SIV-infected rhesus monkeys. *Nature* **540**, 284–287 (2016).
21. Barouch, D. H. et al. Protective efficacy of a global HIV-1 mosaic vaccine against heterologous SHIV challenges in rhesus monkeys. *Cell* **155**, 531–539 (2013).
22. Kawai, T. et al. Interferon- α induction through Toll-like receptors involves a direct interaction of IRF7 with MyD88 and TRAF6. *Nat. Immunol.* **5**, 1061–1068 (2004).
23. Hemmi, H. et al. Small anti-viral compounds activate immune cells via the TLR7 MyD88-dependent signaling pathway. *Nat. Immunol.* **3**, 196–200 (2002).
24. Whitney, J. B. et al. Rapid seeding of the viral reservoir prior to SIV viraemia in rhesus monkeys. *Nature* **512**, 74–77 (2014).
25. Henrich, T. J. et al. Antiretroviral-free HIV-1 remission and viral rebound after allogeneic stem cell transplantation: report of 2 cases. *Ann. Intern. Med.* **161**, 319–327 (2014).
26. Persaud, D. et al. Absence of detectable HIV-1 viremia after treatment cessation in an infant. *N. Engl. J. Med.* **369**, 1828–1835 (2013).
27. Nishimura, Y. et al. Early antibody therapy can induce long-lasting immunity to SHIV. *Nature* **543**, 559–563 (2017).

Acknowledgements We thank K. Reimann and A. Hill for advice, assistance, and reagents. We acknowledge support from the Bill & Melinda Gates Foundation (OPP1107669), the American Foundation for AIDS Research (109219-58-RGRL), the National Institutes of Health (AI096040, AI124377, AI126603, AI129797, AI128751, OD024917), and the Ragon Institute of MGH, MIT, and Harvard.

Author contributions D.H.B. and R.G. designed the study. J.H. and R.G. developed the ART formulation and TLR7 agonist. E.B. conducted the cytokine analyses. E.N.B., J.L., J.P.N., A.M.C., P.A., N.B.M., A.C., D.J., L.P., K.M., and E.T.M. performed the immunologic and virologic assays. W.-H.Y., S.F., T.B., and G.A. led the computational modelling. W.L. led the statistical analysis. M.G.L. led the clinical care of the rhesus monkeys. D.H.B. led the study and wrote the paper with all co-authors.

Competing interests E.B., J.H., and R.G. are employees of Gilead Sciences. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0600-6>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0600-6>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to D.H.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Monkeys and study design. Forty-four outbred Indian-origin, young adult male and female rhesus monkeys (*M. mulatta*) were genotyped, and animals that expressed protective MHC class I alleles and susceptible and resistant TRIM5 α alleles were distributed among the groups. Each group had 0 *Mamu-A*01*, 1 or 2 *Mamu-B*17*, and 0 or 1 *Mamu-B*08* animals. Animals were otherwise randomly allocated to groups. All monkeys were housed at Bioqual, Rockville, MD. Animals were infected with a single 500 TCID₅₀ dose of our rhesus PBMC-derived SHIV-SF162P3 challenge stock²¹ by the intrarectal route. ART was initiated on day 7. Monkeys were bled up to twice per week for viral load determinations. Monkeys received the following interventions starting at week 96 ($n = 11/\text{group}$): (1) sham controls, (2) vesatolimod (GS-9620) alone, (3) PGT121 alone, or (4) both (Extended Data Fig. 1). In groups 2 and 4, animals received 10 oral administrations of 0.15 mg/kg GS-9620 (Gilead Sciences, Foster City, CA) every 2 weeks from weeks 96 to 114. In Groups 3 and 4, animals received 5 intravenous infusions of 10 mg/kg PGT121 (Catalent Biopharma, Madison, WI) every 2 weeks from weeks 106 to 114. Cytokine levels were determined by Luminex assays, and T cell and NK cell activation was assessed by multiparameter flow cytometry. Immunologic and virologic assays were performed blinded. ART was discontinued at week 130. All animal studies complied with all relevant ethical regulations and were approved by the Bioqual Institutional Animal Care and Use Committee (IACUC).

ART regimen. The formulated antiretroviral therapy (ART) cocktail contained 5.1 mg/ml tenofovir disoproxil fumarate (TDF), 40 mg/ml emtricitabine (FTC), and 2.5 mg/ml dolutegravir (DTG) in water containing 15% (v/v) kleptose adjusted to pH 4.2. This ART cocktail was administered once daily at 1 ml/kg body weight via the subcutaneous route.

Adoptive transfer studies. For adoptive transfer studies, 30 million LNMCs and PBMCs from PGT121+GS-9620- or PGT121-treated animals, collected on day 140 following ART discontinuation, were infused intravenously into healthy, SHIV-uninfected rhesus monkeys. Viral loads were assessed in recipient animals weekly following adoptive transfer.

CD8 depletion studies. For CD8 depletion studies, animals received a single intravenous infusion of 50 mg/kg of the anti-CD8 α CDR-grafted rhesus IgG1 antibody MT807R1 (Keith Reimann, MassBiologics, Mattapan, MA). CD8 T cell counts and viral loads were assessed weekly following anti-CD8 infusion.

PGT121 pharmacokinetics. ImmunoClear ELISA plates (Thermo Scientific) were coated with 100 ng/well clade C (C97ZA.012) gp140 Env protein capture reagent overnight at 4°C. Plates were then washed with PBS-0.05% Tween 20 and blocked for 2 h at room temperature with Blocker Casein in PBS (Pierce). Standard curve calibrators and diluted serum samples were incubated on the plates for 1 h before further washing and subsequent incubation with a mouse PGT121 anti-idiotype (1 $\mu\text{g}/\text{ml}$) monoclonal antibody. Plates were washed again and then incubated with 1:1,000 dilution of rabbit anti-mouse IgG-horseradish peroxidase (Thermo Scientific). Finally, plates were washed and developed for 5 min using SureBlue (KPL Laboratories) followed by addition of TMB Stop solution (KPL Laboratories). Plates were read at 450 nm on a VersaMax microplate reader (Molecular Devices) using Softmax Pro version 6.5.1 software. The SoftMax Pro software calculated 4-Parameter Logistic (4-PL) curve fits for the standard calibrators and the test sample concentrations were determined by interpolation into the calibration curves.

PGT121 anti-drug antibody (ADA) assays. ADA assays were performed using the MesoScale Discovery (MSD) electrochemiluminescence (ECL) platform. In brief, serum samples were diluted in assay buffer and incubated with a master mix containing sulfo-tagged and biotinylated-tagged PGT121 (at equimolar concentrations of 0.5 $\mu\text{g}/\text{ml}$) overnight at 4°C on an orbital plate shaker. The mixture was then incubated on a streptavidin-functionalized plate and washed, and tripropylamine (TPA) was added and read using an MSD SQ-120 ECL imager. Luminescence was proportional to the amount of ADA. The assay positivity cut-point was defined as the 95th percentile of ECL signal response observed using naive NHP serum samples.

Cellular immune assays. SIV-specific cellular immune responses were assessed using IFN- γ ELISPOT assays and multiparameter intracellular cytokine staining (ICS) assays essentially as described²⁰. ICS assays were performed with 10 6 PBMCs or LNMCs that were incubated for 6 h at 37°C with medium, 10 pg/ml phorbol myristate acetate (PMA) and 1 $\mu\text{g}/\text{ml}$ ionomycin (Sigma-Aldrich), or 1 $\mu\text{g}/\text{ml}$ HIV-1 Env, SIV Gag, or SIV Pol peptide pools. Cultures contained monensin (GolgiStop; BD Biosciences), brefeldin A (GolgiPlug; BD Biosciences), and 1 $\mu\text{g}/\text{ml}$ of a mAb against human CD49d (clone 9F10). Cells were then stained with predetermined titres of mAbs against CD3 (clone SP34.2; Alexa 700), CD4 (clone L200; BV786), CD8 (clone SK1; APC H7), CD28 (clone L293, PerCP-Cy5.5), CD95 (clone DX2, BV711), CD20 (clone 2H10, BV570), CCR5 (clone 3A9, PE), BCL6 (clone K112-91, PE-CF594), CXCR5 (clone MU5UBEE, PE-cy7), and PD-1 (clone EH12.2H7, Pacific Blue); and stained intracellularly with IFN- γ (clone B27; BUV395), IL-2 (clone MQ1-17H12; BUV737), TNF- α (clone Mab11; BV650),

CD69 (clone FN50, BV510), and Ki67 (clone B56, FITC). IFN- γ backgrounds were <0.05% in PBMCs.

Viral RNA assays. Viral RNA was isolated from cell-free plasma using a viral RNA extraction kit (Qiagen) and was quantified essentially as described²⁴.

Viral DNA assays. Levels of proviral DNA were quantified as previously described²⁴. Total cellular DNA was isolated from 5 $\times 10^6$ cells using a QIAamp DNA Blood Mini kit (Qiagen). The absolute quantification of viral DNA in each sample was determined by qPCR using primers specific to a conserved region SIVmac239. All samples were directly compared to a linear virus standard and the simultaneous amplification of a fragment of human GAPDH gene. PCR assays were performed with 100–200 ng sample DNA.

Computational modelling. The frequencies of cell populations at week 106 (corresponding to the sixth GS-9620 administration) before and after the intervention were used for the computational model. Z-score standardization was performed to have all features mean centred and unit variance scaled. The minimal correlates that best predicted viral rebound (or total viral loads) were identified by using a two-step model: least absolute shrinkage and selection operator (LASSO) regularization followed by partial least squares regression analysis (PLSR). The LASSO method was used to remove irrelevant features in order to improve the robustness of high-dimensional modelling. Next, PLSR was used to model the covariance relationship between the feature variables (X) and the outcome variables (Y), in the way that PLSR decomposes both X and Y into a hyperplane and maximizes covariance between the hyperplanes. To estimate the minimal correlates that best explain the outcome without overfitting, 5,000 repeated fivefold nested cross-validation was designed. In each repetition, the data set was randomly divided into five folds, where 80% of the data set was used for building the model and the remaining holdout set was used to test the model prediction, where the goodness-of-fit of the model was measured by mean squared error (MSE) between prediction and the outcome. This approach resulted in the generation of a model with the minimal set of the features that generates the best outcome prediction in cross-validation test. In addition, variable importance in projection (VIP), a weighted sum of squares of the PLSR weights that summarized the importance of the features in a PLSR model with multiple components, was computed. To estimate the statistical significance of the optimized model with the defined correlates, we employed two types of permutation tests (shuffling the outcome label and selecting the randomized correlates) to test the likelihood of obtaining a model prediction accuracy by chance. Each permutation test performed 1,000 times to generate an empirical null distribution, and an exact P value of the model was computed.

PGT121-mediated CD4 $^+$ T cell killing assay. Human PBMCs were isolated from fresh blood from healthy donors using SepMate PBMC Isolation tubes (StemCell Technologies) and histopaque (Sigma Aldrich) according to the manufacturer's protocol. Using an aliquot of PBMCs, CD4 $^+$ T cells were isolated for HIV-1 infection using the EasySep Human CD4 $^+$ T Cell Isolation Kit (StemCell Technologies). Remaining PBMCs were treated with 1,000 nM GS-9620 or DMSO and cultured at 3.0 $\times 10^6$ cells/ml for 5 days in RPMI 1640 (Sigma Aldrich) medium supplemented with L-glutamine, HEPES, and IL-15 (1 ng/ml) at 37°C. For HIV-1 infection, freshly isolated CD4 $^+$ T cells were spinfected for 45 min at 1,500 RPM with NL4-3 virus (or medium for mock infection) and then treated with CD3/CD28 Human T Activator Dynabeads (ThermoFisher Scientific). Infected CD4 $^+$ T cells were subsequently cultured for 5 days in R10 and IL-2 (30 units/ml) at 37°C. Cells were counted daily and supplemented with fresh medium with IL-2 as needed. On day 5, both sets of cells were counted, washed twice and resuspended in fresh medium. HIV-infected CD4 $^+$ T cells were incubated with 10 $\mu\text{g}/\text{ml}$ of PGT121 for 30 min. Co-cultures of infected CD4 $^+$ T cells and autologous donor PBMCs were then set up at a ratio of 1:10 (5.0 $\times 10^4$ CD4 $^+$ T cells; 5.0 $\times 10^5$ PBMCs) and incubated overnight in R10 and IL-2 at 37°C. Following the overnight incubation, cells were stained with Fixable Aqua viability dye (Invitrogen) followed by surface staining with CD4-APC (Invitrogen), CD56-PE-Cy7 (BD), CD3-AF700 (BD), and CD69-BV605 (BD). Cells were fixed and permeabilized with Cytofix/Cytoperm (BD Biosciences) and stained intracellularly with anti-KC57-PE-labelled p24 antibody (Beckman Coulter) in 1 \times Perm/Wash Buffer (BD Biosciences, 554723). Gating on viable CD3 $^+$ CD4 $^+$ p24 $^+$ T cells was used to evaluate viral killing with each treatment condition normalized to an uninfected control. Per cent killing was normalized to the untreated antibody condition.

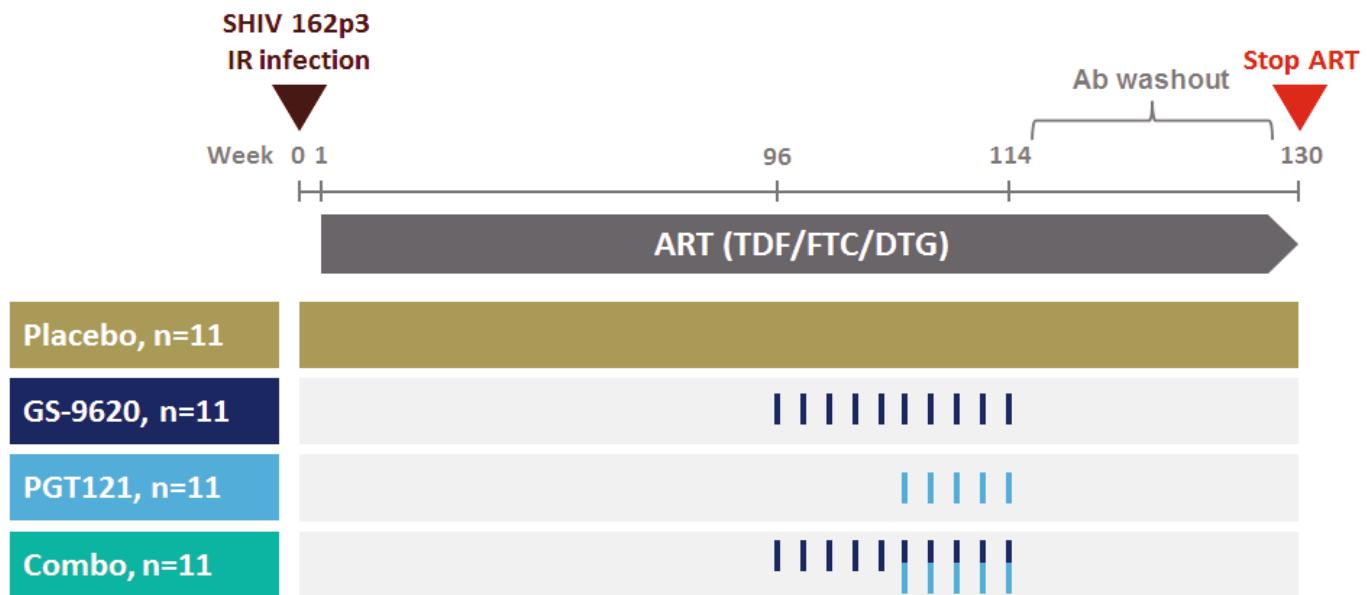
Statistical analyses. Analysis of virologic and immunologic data was performed using GraphPad Prism v6.03 (GraphPad Software). Comparisons of groups was performed using two-sided Mann-Whitney tests without Bonferroni adjustments. Correlations were assessed by two-sided Spearman rank-correlation tests. Analysis of viral rebound was performed using Kruskall-Wallis tests to compare all groups. For group pairwise comparisons, area under the curve (AUC) for total viral RNA following ART discontinuation was compared with chi-square tests, and viral rebound was compared with a censored Poisson regression model. In vitro killing data were analysed with paired Student's t -tests.

No statistical methods were used to predetermine sample size. Except as stated, the experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

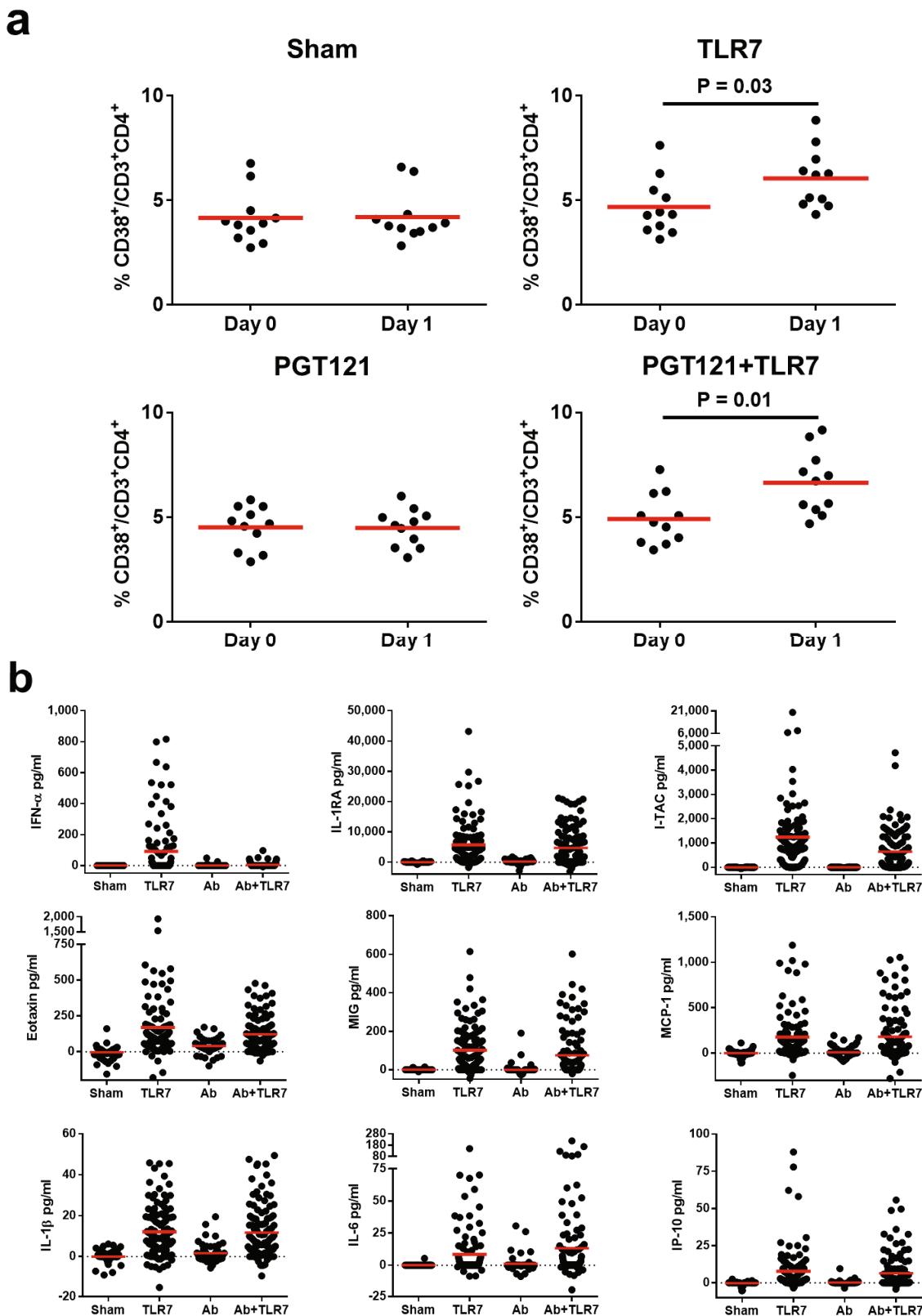
Data availability

All data generated and analysed in this study are available from the corresponding author upon reasonable request. Source data for figures from individual animals are available online.



Extended Data Fig. 1 | Study design. Forty-four rhesus monkeys ($n=11$ monkeys per group) were infected with SHIV-SF162P3 at week 0 and ART was initiated at week 1 (day 7). GS-9620 administrations and PGT121

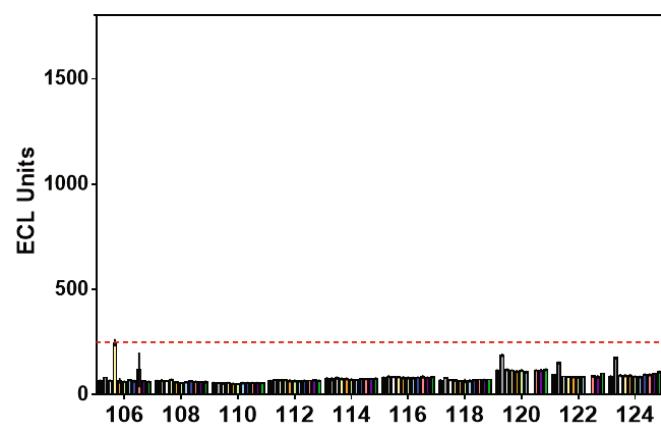
infusions are shown from weeks 96 to 114. ART was discontinued at week 130.



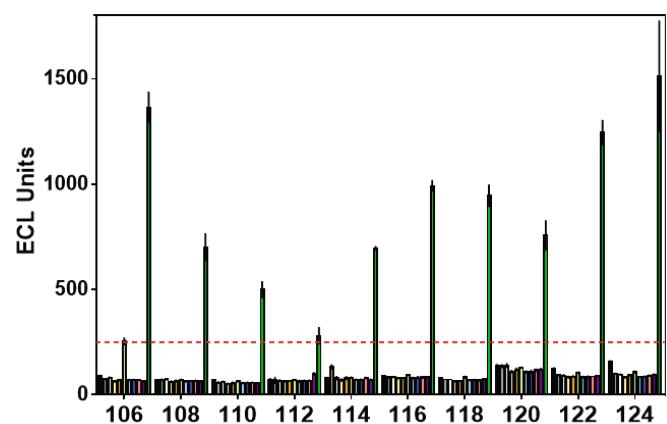
Extended Data Fig. 2 | Immune activation following GS-9620 administration and before ART discontinuation. **a**, Activation of CD4 $^{+}$ T cells was assessed by CD38 expression on days 0 and 1 following GS-9620 administration, supplementing the data shown in Fig. 2a ($n = 11$ monkeys per group). Representative data are shown following the fifth GS-9620 dose, which was comparable to the other doses. Red horizontal

bars indicate median values. P values reflect two-sided Mann–Whitney tests. **b**, Plasma levels of IFN- α , IL-1RA, I-TAC, eotaxin, MIG, MCP-1, IL-1 β , IL-6, IP-10 are shown on day 1 following GS-9620 administration ($n = 11$ monkeys per group). Red bars represent mean values. Combined data from all GS-9620 administrations with pre-dose levels subtracted are shown.

PGT121



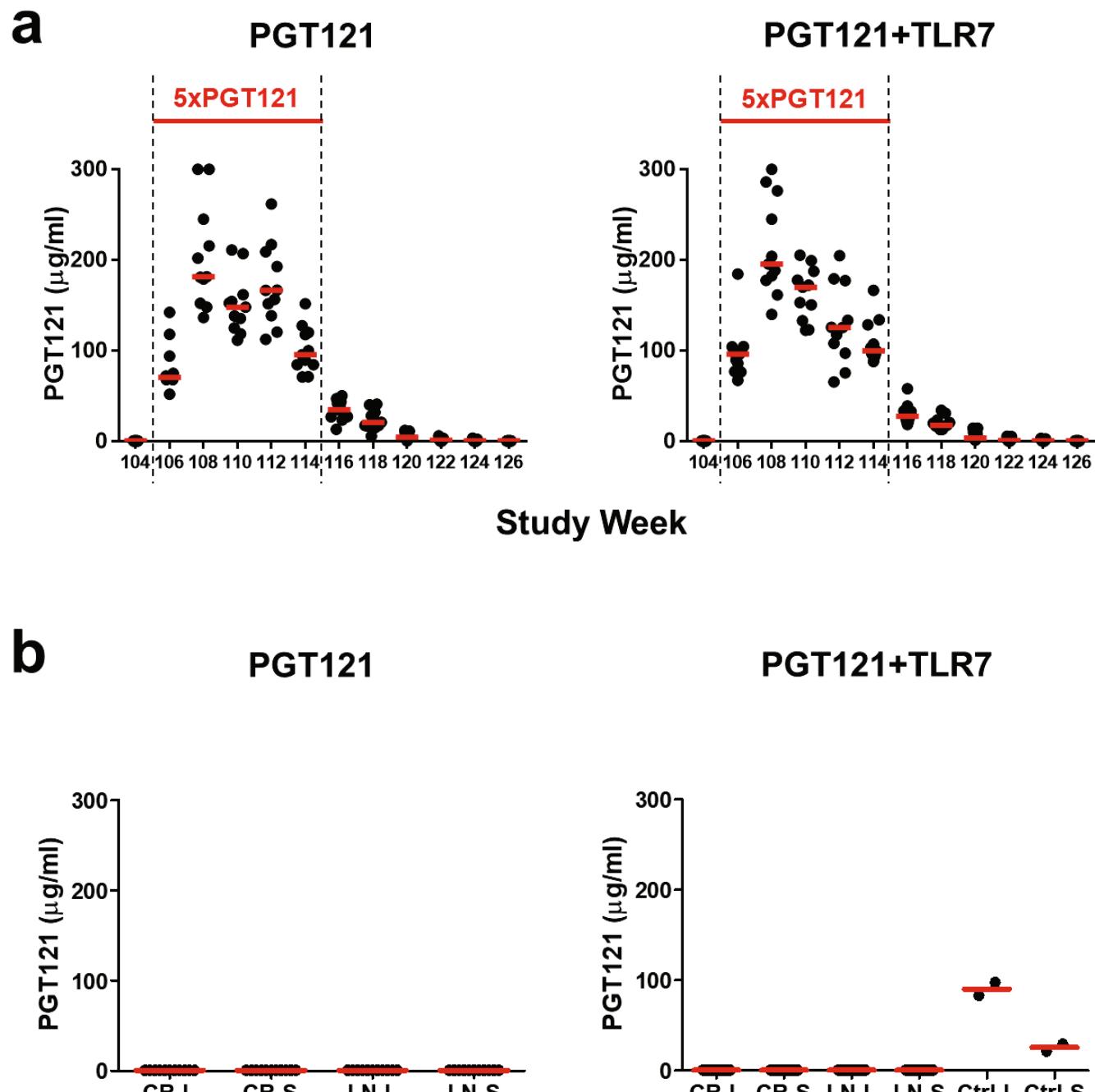
PGT121+TLR7



Study Week

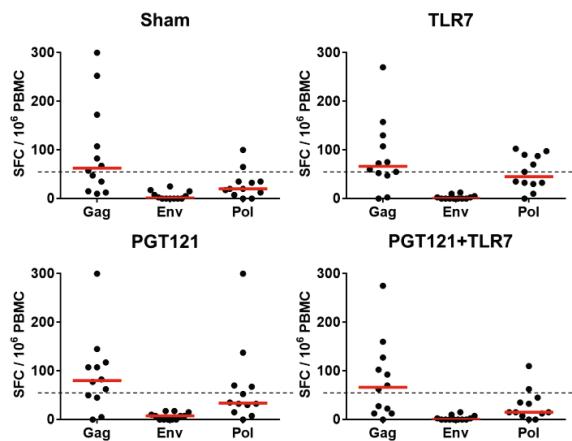
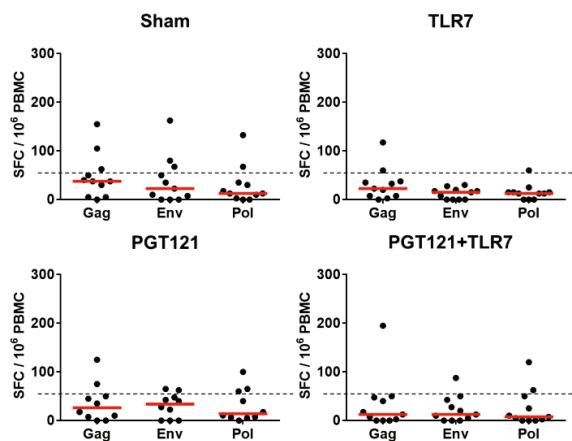
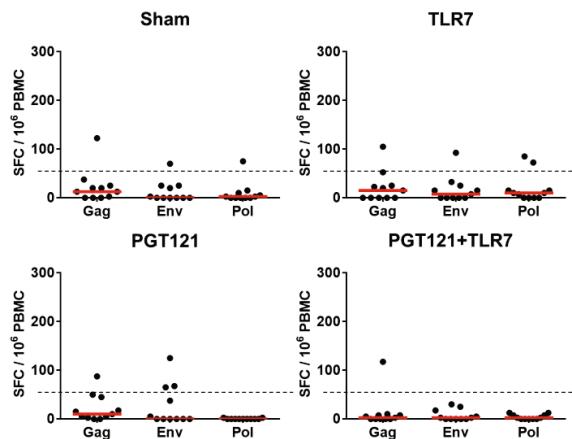
Extended Data Fig. 3 | Anti-drug antibody (ADA) assay before ART discontinuation. ADA responses were assessed in the PGT121+GS-9620 and PGT121-alone groups every 2 weeks from weeks 106 to 124 using an electrochemiluminescence (ECL) assay with an anti-PGT121 idiotypic

mAb ($n = 11$ monkeys per group). No ADA was detected. One monkey in the PGT121+GS-9620 group had background reactivity in this assay at week 106 before PGT121 exposure (green bars).

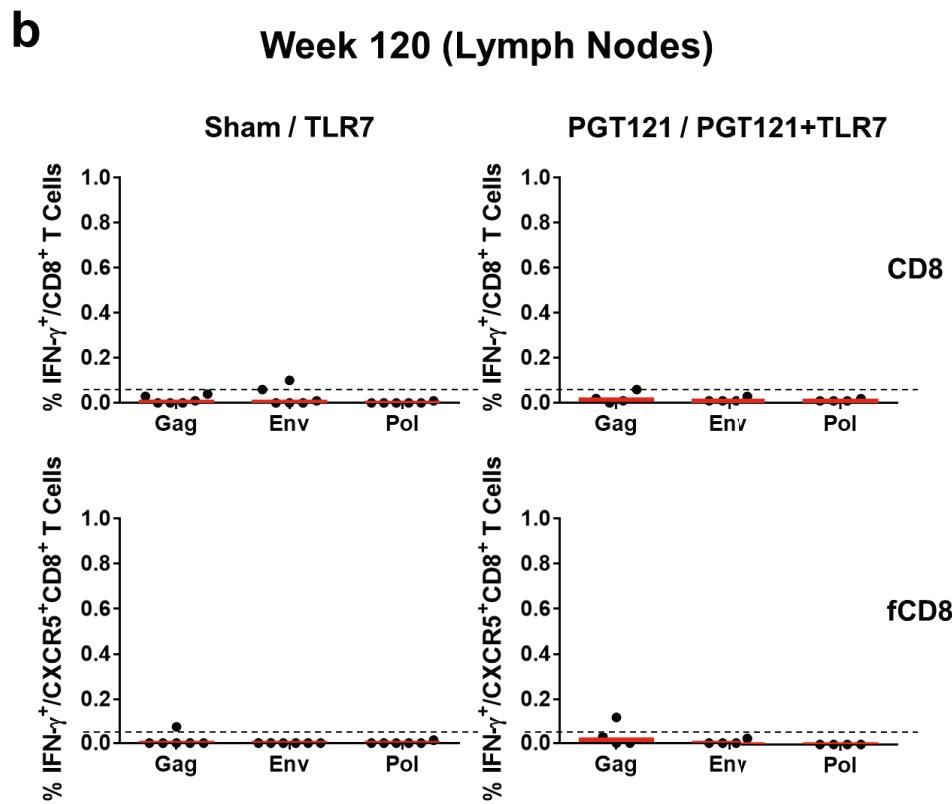
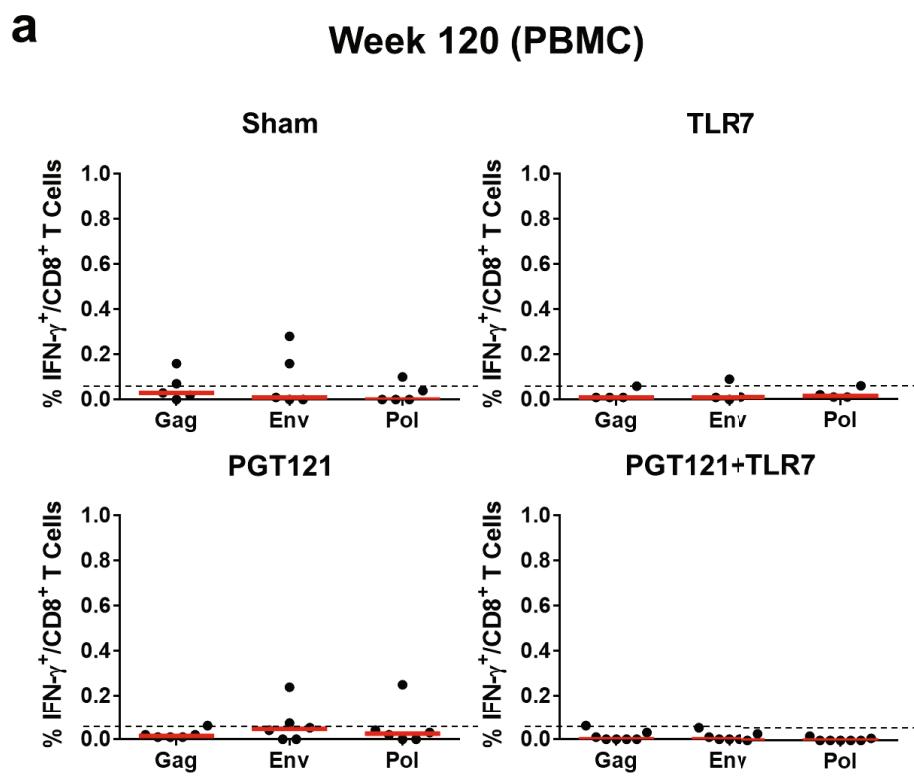


Extended Data Fig. 4 | PGT121 pharmacokinetics in serum and tissues before ART discontinuation. **a**, Peak serum PGT121 levels are shown (limit of detection $0.5 \mu\text{g ml}^{-1}$) 1 h following each of five infusions of PGT121 (weeks 106–114) and during the washout period (weeks 114–130) ($n = 11$ monkeys per group). **b**, PGT121 levels (limit of detection

$0.5 \mu\text{g ml}^{-1}$) were assessed in cell lysates (L) and initial wash supernatants (S) from 10^6 lymph node (LN) and colorectal (CR) cells from week 120 ($n = 11$ monkeys per group). Positive controls (Ctrl) included lymph node samples from naive monkeys spiked with PGT121. Red bars represent median values.

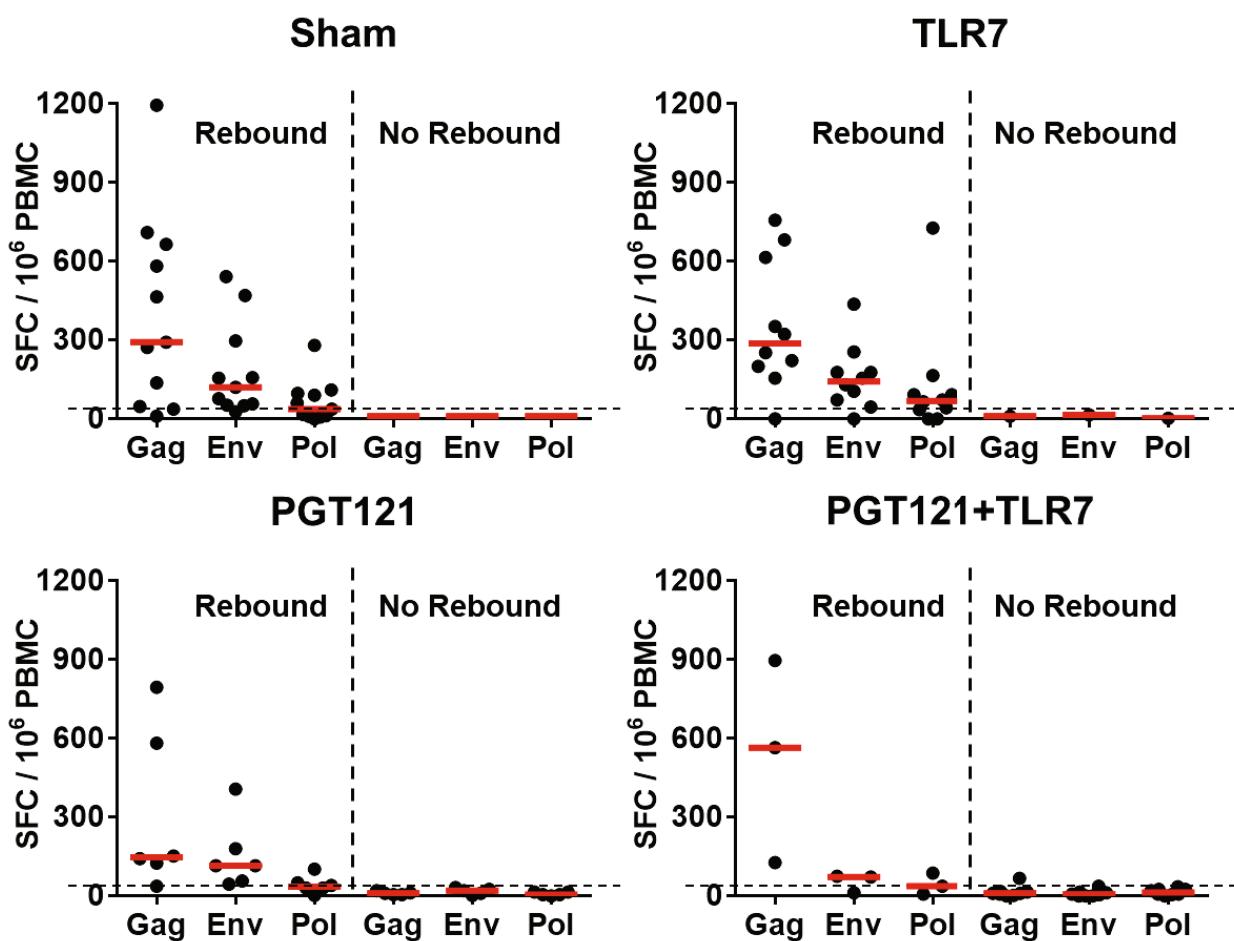
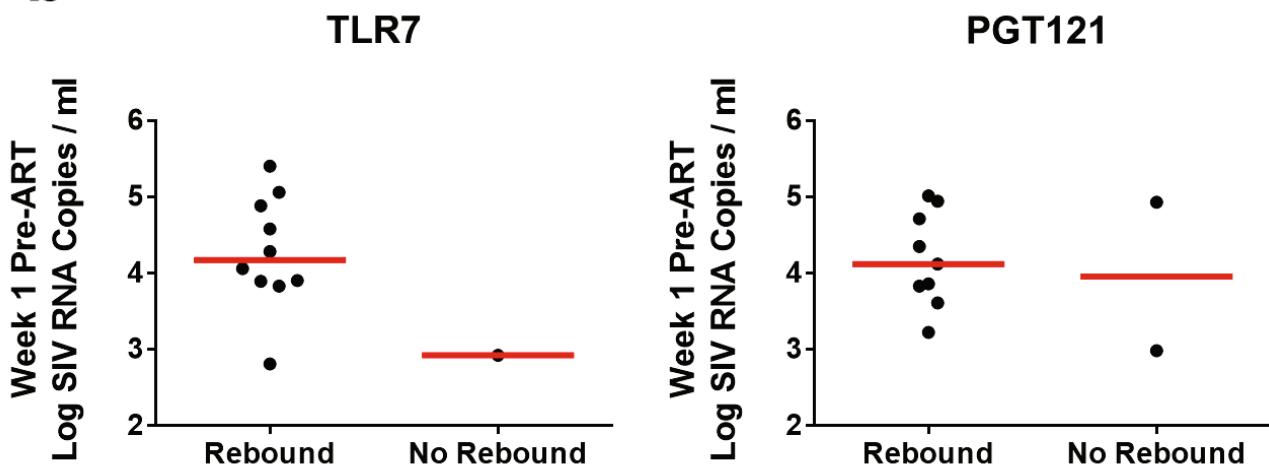
a**Week 4 (PBMC)****b****Week 96 (PBMC)****c****Week 120 (PBMC)**

Extended Data Fig. 5 | IFN γ ELISPOT responses before ART discontinuation. Gag-, Env-, and Pol-specific IFN γ ELISPOT responses in PBMCs are shown at week 4 (a), week 96 (b), and week 120 (c) ($n=11$ monkeys per group). Spot-forming cells (SFCs) per million PBMCs are shown. Red horizontal bars indicate median values. The dotted line represents the assay limit of quantitation (55 SFCs per million PBMCs).



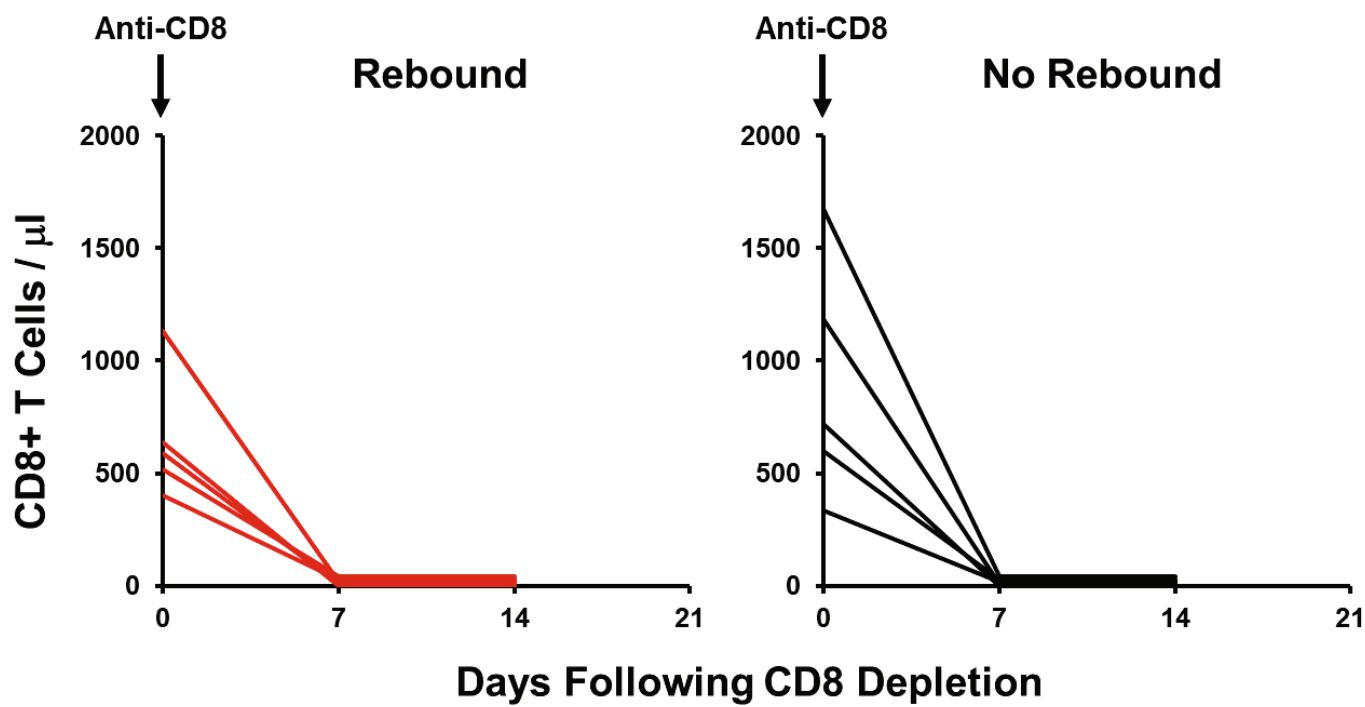
Extended Data Fig. 6 | IFN γ intracellular cytokine staining (ICS) responses before ART discontinuation. Gag-, Env-, and Pol-specific IFN γ ICS responses in PBMCs (a) and in LNMCs (b) are shown at week 120 ($n=11$ monkeys per group). Per cent IFN γ -producing CD8 $^{+}$ CD3 $^{+}$

T cells in PBMCs and per cent IFN γ -producing total CD8 $^{+}$ CD3 $^{+}$ T cells (CD8) and follicular CXCR5 $^{+}$ CD8 $^{+}$ CD3 $^{+}$ T cells (fCD8) in LNMCs are shown. Red horizontal bars indicate median values. The dotted line represents the assay limit of quantitation (0.05% CD8 $^{+}$ CD3 $^{+}$ T cells).

a**b**

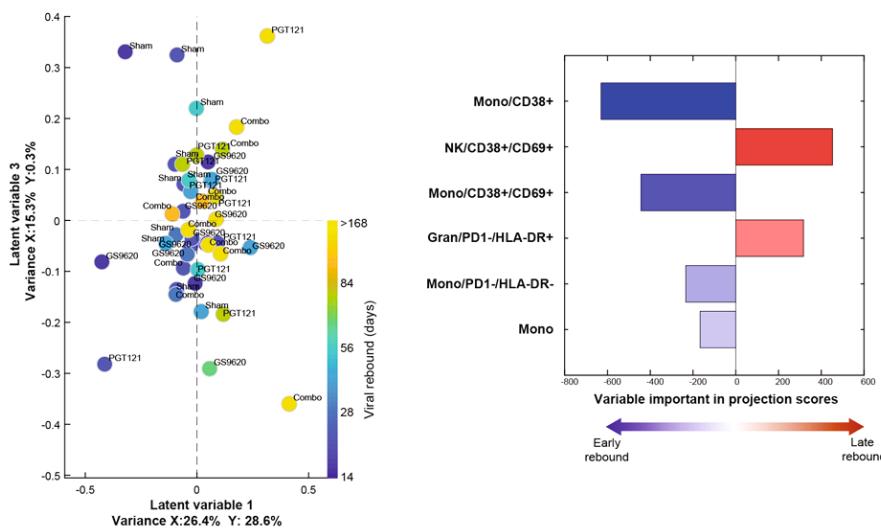
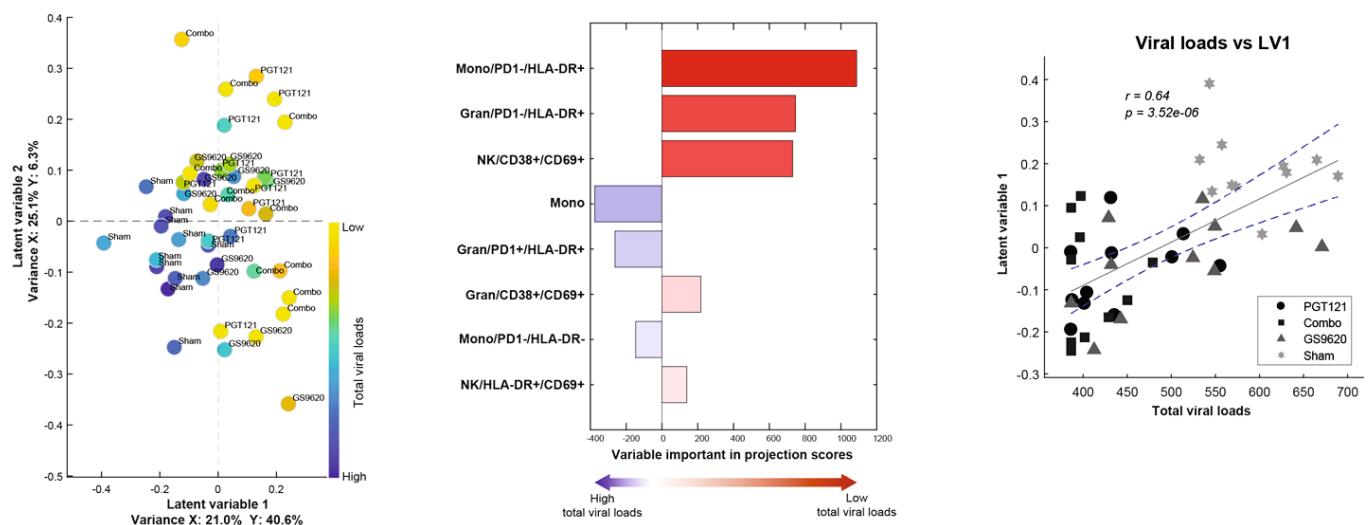
Extended Data Fig. 7 | IFN γ ELISpot responses following ART discontinuation and trends for viral rebound. a, Gag-, Env-, and Pol-specific IFN γ ELISpot responses in PBMCs are shown at day 140 following ART discontinuation ($n=11$ monkeys per group). SFCs per million PBMCs are shown. Monkeys in each group that demonstrated viral rebound versus no rebound are shown separately. Red horizontal

bars indicate median values. The dotted line represents the assay limit of quantitation (55 SFCs per million PBMCs). **b,** Trends for viral rebound are shown in the GS-9620 and PGT121 groups in relation to pre-ART week 1 viral loads, supplementing the data shown in Fig. 5c. Red horizontal bars indicate median values.



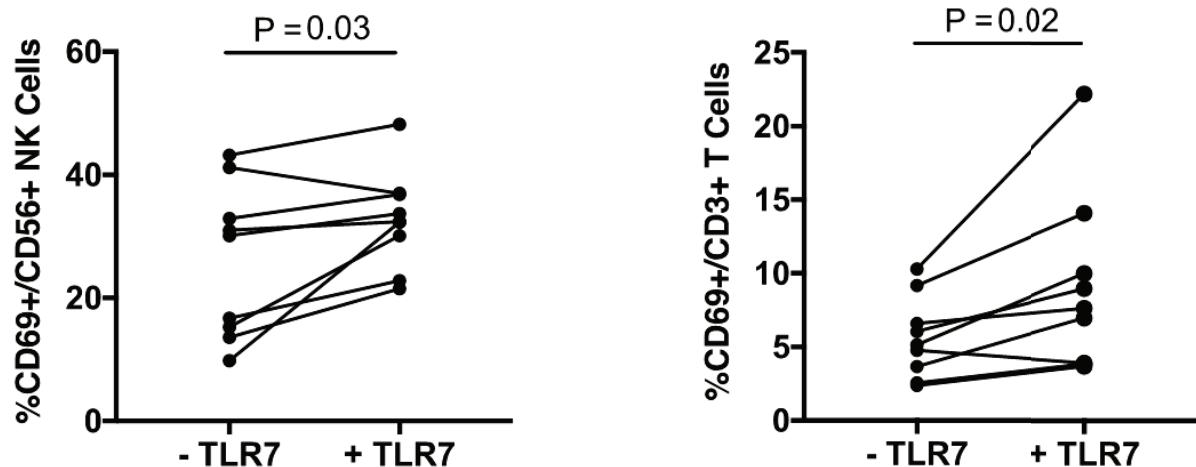
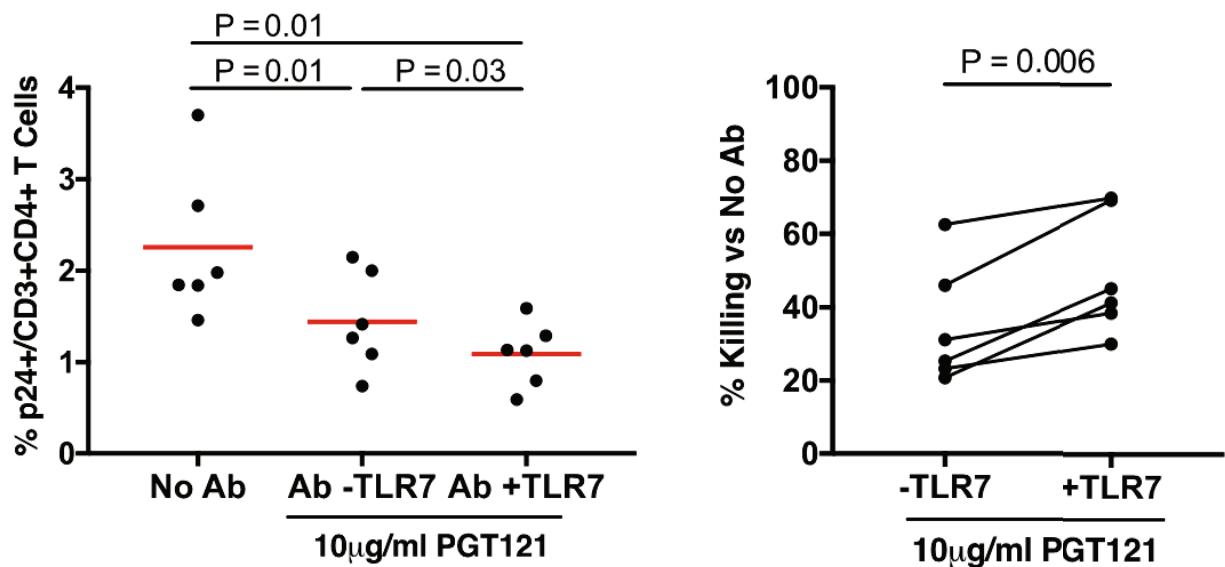
Extended Data Fig. 8 | CD8 depletion efficiency. CD8⁺ T cells per μ l peripheral blood are shown in PGT121+GS-9620-treated monkeys before and after CD8 depletion in monkeys that exhibited viral rebound and

post-rebound virologic control ($n = 5$, left, red lines) and in monkeys that exhibited no viral rebound following ART discontinuation ($n = 5$, right, black lines).

a**b**

Extended Data Fig. 9 | Computational model. a, LASSO and PLSR model identifies the parameters that correlate with delayed viral rebound ($n = 11$ monkeys per group). Left, individual monkeys are shown distributed by latent variables 1 and 3 of the model. Timing of viral rebound is indicated by the colour gradient. $R^2 = 0.176$, root mean square error (RMSE) = 0.917, $P < 0.001$ in two-sided permutation tests. Middle, the contribution of the selected features to model separation is displayed in variable importance in projection (VIP) scores, related to early (blue) or late (red) viral rebound. Right, correlation between viral rebound and latent variable 1. P value reflects a two-sided Spearman rank-correlation

test. **b**, LASSO and PLSR model identifies the parameters that correlate with reduced total viral loads ($n = 11$ monkeys per group). Left, individual monkeys are showed distributed by latent variables 1 and 2. Total viral loads are indicated by the colour gradient. $R^2 = 0.282$, root mean square error (RMSE) = 0.857, $P < 0.001$ in two-sided permutation tests. Middle, the contribution of the selected features to model separation is displayed in VIP scores, related to high (blue) or low (red) total viral loads. Right, correlation between total viral loads and latent variable 1. P value reflects a two-sided Spearman rank-correlation test.

a**b**

Extended Data Fig. 10 | In vitro killing studies. **a**, GS-9620 treatment led to CD69 upregulation of CD56⁺ NK cells and CD3⁺ T cells in vitro following incubation of human PBMCs with 1,000 nM GS-9620 for 5 days ($n=9$). **b**, GS-9620 treatment augmented PGT121-mediated killing of PGT121 in vitro. Per cent p24 reduction in CD4⁺ T cells ($n=6$) using

an antibody-mediated killing assay (see Methods). Per cent killing was calculated as the per cent reduction in p24 in CD4⁺ T cells with PGT121 compared with no PGT121. P values reflect two-sided paired Student's t -tests.

A candidate super-Earth planet orbiting near the snow line of Barnard's star

I. Ribas^{1,2*}, M. Tuomi³, A. Reiners⁴, R. P. Butler⁵, J. C. Morales^{1,2}, M. Perger^{1,2}, S. Dreizler⁴, C. Rodríguez-López⁶, J. I. González Hernández^{7,8}, A. Rosich^{1,2}, F. Feng³, T. Trifonov⁹, S. S. Vogt¹⁰, J. A. Caballero¹¹, A. Hatzes¹², E. Herrero^{1,2}, S. V. Jeffers⁴, M. Lafarga^{1,2}, F. Murgas^{7,8}, R. P. Nelson¹³, E. Rodríguez⁶, J. B. P. Strachan¹³, L. Tal-Or^{4,14}, J. Teske⁵, B. Toledo-Padrón^{7,8}, M. Zechmeister⁴, A. Quirrenbach¹⁵, P. J. Amado⁶, M. Azzaro¹⁶, V. J. S. Béjar^{7,8}, J. R. Barnes¹⁷, Z. M. Berdiñas¹⁸, J. Burt¹⁹, G. Coleman²⁰, M. Cortés-Contreras¹¹, J. Crane²¹, S. G. Engle²², E. F. Guinan²², C. A. Haswell¹⁷, Th. Henning⁹, B. Holden¹⁰, J. Jenkins¹⁸, H. R. A. Jones³, A. Kaminski¹⁵, M. Kiraga²³, M. Kürster⁹, M. H. Lee²⁴, M. J. López-González⁶, D. Montes²⁵, J. Morin²⁶, A. Ofir²⁷, E. Pallé^{7,8}, R. Rebolo^{7,8,28}, S. Reffert¹⁵, A. Schweitzer²⁹, W. Seifert¹⁵, S. A. Shectman²¹, D. Staab¹⁷, R. A. Street³⁰, A. Suárez Mascareño^{7,31}, Y. Tsapras³², S. X. Wang⁵ & G. Anglada-Escudé^{6,13}

Barnard's star is a red dwarf, and has the largest proper motion (apparent motion across the sky) of all known stars. At a distance of 1.8 parsecs¹, it is the closest single star to the Sun; only the three stars in the α Centauri system are closer. Barnard's star is also among the least magnetically active red dwarfs known^{2,3} and has an estimated age older than the Solar System. Its properties make it a prime target for planetary searches; various techniques with different sensitivity limits have been used previously, including radial-velocity imaging^{4–6}, astrometry^{7,8} and direct imaging⁹, but all ultimately led to negative or null results. Here we combine numerous measurements from high-precision radial-velocity instruments, revealing the presence of a low-amplitude periodic signal with a period of 233 days. Independent photometric and spectroscopic monitoring, as well as an analysis of instrumental systematic effects, suggest that this signal is best explained as arising from a planetary companion. The candidate planet around Barnard's star is a cold super-Earth, with a minimum mass of 3.2 times that of Earth, orbiting near its snow line (the minimum distance from the star at which volatile compounds could condense). The combination of all radial-velocity datasets spanning 20 years of measurements additionally reveals a long-term modulation that could arise from a stellar magnetic-activity cycle or from a more distant planetary object. Because of its proximity to the Sun, the candidate planet has a maximum angular separation of 220 milliarcseconds from Barnard's star, making it an excellent target for direct imaging and astrometric observations in the future.

Barnard's star is the second closest red dwarf to the Solar System, after Proxima Centauri, and therefore an ideal target for searches for exoplanets that have the potential for further characterization¹⁰. Its very low X-ray flux, lack of H α emission, low chromospheric emission indices, slow rotation rate, slightly subsolar metallicity and membership of the thick-disk kinematic population are indicative of extremely low magnetic activity and an age older than the Sun. Because of its apparent brightness and very low variability, Barnard's star is often regarded as

a benchmark for intermediate M-type dwarfs. Its basic properties are summarized in Table 1.

An early analysis of archival radial-velocity datasets of Barnard's star up to 2015 indicated the presence of at least one significant signal, which had a period of about 230 days, but with rather poor sampling. To elucidate its presence and nature we undertook an intensive monitoring campaign with the CARMENES spectrometer¹¹, collecting precise radial-velocity measurements on every possible night during 2016 and 2017. We also obtained overlapping observations with the European Southern Observatory (ESO) HARPS and the HARPS-N instruments. The combined Doppler monitoring of Barnard's star, including archival and newly acquired observations, resulted in 771 radial-velocity epochs (nightly averages), with typical individual precisions of 0.9–1.8 m s $^{-1}$, obtained over a timespan of more than 20 years from seven different facilities, and yielded eight independent datasets (Extended Data Table 1).

Although each dataset is internally consistent, relative offsets may be present because of uncertainties in the absolute radial-velocity scale. Our analysis considers a zero-point value and a noise term (jitter) for each dataset as free parameters to be optimized simultaneously with the planetary models, and a global linear trend. We used several independent fitting methods to ensure the reliability of the results. The parameter space was scanned using hierarchical procedures (signals are identified individually and added recursively to the model) and multi-signal search approaches (fitting two or more signals at a time). Furthermore, we used the Systemic Console¹² to assess the sensitivity of the solutions to the datasets used, to the error estimates and to the eccentricity. Figure 1 and Extended Data Fig. 1 illustrate the detection of a signal with a period of 233 days with high statistical significance from an analysis assuming uncorrelated (white) noise (P value or false-alarm probability (FAP) of roughly 10^{-15}) and show evidence for a second, longer-period signal.

To assess the presence of the long-term modulation we considered an alternative method of determining the relative offsets, which involves

¹Institut de Ciències de l'Espai (ICE, CSIC), Campus UAB, Bellaterra, Spain. ²Institut d'Estudis Espacials de Catalunya (IEEC), Barcelona, Spain. ³Centre for Astrophysics Research, University of Hertfordshire, Hatfield, UK. ⁴Institut für Astrophysik Göttingen, Georg-August-Universität Göttingen, Göttingen, Germany. ⁵Department of Terrestrial Magnetism, Carnegie Institution for Science, Washington, DC, USA. ⁶Instituto de Astrofísica de Andalucía (IAA, CSIC), Granada, Spain. ⁷Instituto de Astrofísica de Canarias (IAC), La Laguna, Spain. ⁸Universidad de La Laguna (ULL), Departamento de Astrofísica, La Laguna, Spain. ⁹Max-Planck-Institut für Astronomie, Heidelberg, Germany. ¹⁰UCO/Lick Observatory, University of California at Santa Cruz, Santa Cruz, CA, USA. ¹¹Centro de Astrobiología, CSIC-INTA, ESAC, Villanueva de la Cañada, Spain. ¹²Thüringer Landessternwarte, Tautenburg, Germany. ¹³School of Physics and Astronomy, Queen Mary University of London, London, UK. ¹⁴School of Geosciences, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel-Aviv University, Tel Aviv, Israel. ¹⁵Landessternwarte, Zentrum für Astronomie der Universität Heidelberg, Heidelberg, Germany. ¹⁶Centro Astronómico Hispano-Alemán (CSIC-MPG), Observatorio Astronómico de Calar Alto, Gérgal, Spain. ¹⁷School of Physical Sciences, The Open University, Milton Keynes, UK. ¹⁸Departamento de Astronomía, Universidad de Chile, Santiago, Chile. ¹⁹Kavli Institute, Massachusetts Institute of Technology, Cambridge, MA, USA. ²⁰Physikalisches Institut, Universität Bern, Bern, Switzerland. ²¹The Observatories, Carnegie Institution for Science, Pasadena, CA, USA. ²²Department of Astrophysics and Planetary Science, Villanova University, Villanova, PA, USA. ²³Warsaw University Observatory, Warsaw, Poland. ²⁴Department of Earth Sciences and Department of Physics, The University of Hong Kong, Pok Fu Lam, Hong Kong. ²⁵Departamento de Física de la Tierra Astronomía y Astrofísica and UPARCOS-UCM (Unidad de Física de Partículas y del Cosmos de la UCM), Facultad de Ciencias Físicas, Universidad Complutense de Madrid, Madrid, Spain. ²⁶Laboratoire Univers et Particules de Montpellier, Université de Montpellier, CNRS, Montpellier, France. ²⁷Department of Earth and Planetary Sciences, Weizmann Institute of Science, Rehovot, Israel. ²⁸Consejo Superior de Investigaciones Científicas (CSIC), Madrid, Spain. ²⁹Hamburger Sternwarte, Universität Hamburg, Hamburg, Germany.

³⁰Las Cumbres Observatory Global Telescope Network, Goleta, CA, USA. ³¹Observatoire Astronomique de l'Université de Genève, Versoix, Switzerland. ³²Zentrum für Astronomie der Universität Heidelberg, Astronomisches Rechen-Institut, Heidelberg, Germany. *e-mail: iribas@ice.cat

Table 1 | Information on Barnard's star and its planet

Stellar parameter	Value
Spectral type	M3.5 V
Mass (M_{\odot})	0.163 ± 0.022
Radius (R_{\odot})	0.178 ± 0.011
Luminosity (L_{\odot})	0.00329 ± 0.00019
Effective temperature (K)	$3,278 \pm 51$
Rotation period (d)	140 ± 10
Age (Gyr)	7–10
Planetary parameter	Value
Orbital period (d)	$232.80^{+0.38}_{-0.41}$
Radial-velocity semi-amplitude ($m s^{-1}$)	1.20 ± 0.12
Eccentricity	$0.32^{+0.10}_{-0.15}$
Argument of periastron (°)	107^{+19}_{-22}
Mean longitude at BJD 2,455,000.0 (°)	203 ± 7
Minimum mass, $M_{\text{sin}i}$ (M_{\oplus})	3.23 ± 0.44
Orbital semi-major axis (AU)	0.404 ± 0.018
Irradiance (Earth units)	0.0203 ± 0.0023
Maximum equilibrium temperature (K)	105 ± 3
Minimum astrometric semi-amplitude, $\alpha_{\text{sin}i}$ (mas)	0.0133 ± 0.0013
Angular separation (mas)	221 ± 10

We derive fundamental parameters of Barnard's star as in ref. ²⁹. The luminosity is calculated from a well-sampled spectral energy distribution and the effective temperature is used to derive the stellar radius. The age interval is estimated by considering kinematic parameters, stellar rotation and indicators of magnetic activity. The planetary parameters and their uncertainties are determined as the median values and 68% credibility intervals of the distribution that results from the MCMC run. The maximum equilibrium temperature is calculated assuming only external energy sources and a null Bond albedo. M_{\odot} , R_{\odot} and L_{\odot} are the mass, radius and luminosity of the Sun; M_{\oplus} is the mass of Earth; i is the orbital inclination; M is the true planetary mass; α is the true astrometric semi-amplitude; BJD, barycentric Julian date.

directly averaging radial-velocity differences within defined time intervals for overlapping observations. All datasets were subsequently stitched together into a single radial-velocity time series. These combined measurements indicate long-term variability consistent with a signal with a period of more than 6,000 days. We therefore performed additional fits leaving the relative offsets as free parameters and assuming two signals, one with a prior allowing only periods of more than 4,000 days. The model fit converges to two periodic signals at 233 days and about 6,600 days and has comparable likelihood ($\Delta \ln L < 5$) to that obtained by manually stitching the datasets. We conclude that the significance of the 233-day signal remains unaltered irrespective of the model used for the long-term variability and that the long-term variability is significant.

Stellar activity is known to produce periodic radial-velocity modulations that could be misinterpreted as arising from planetary companions. Rotation periods of 130 days and 148.6 days have been reported for Barnard's star from photometry¹³ and from spectroscopic indices³, respectively. We analysed data from long-term monitoring in photometry and spectroscopy, the latter being H α and Ca II H + K chromospheric fluxes measured from the spectra used for radial-velocity determination. Periodograms are shown in Fig. 2. The photometric time series yields a statistically significant signal with a period of 144 days, the H α measurements present a complex periodogram with a highly significant main peak at 133 days and the Ca II H + K chromospheric index shows significant periodicity at 143 days. All of these values can be tentatively associated with the stellar rotation period, which we estimate to be 140 ± 10 days. Furthermore, two of the activity tracers suggest the existence of long-term variability. The analysis rules out stellar-activity periodicities in the neighbourhood of 230 days. Also, the significance of the 233-day signal in radial velocity increases mostly monotonically with time as additional observations are accumulated (Extended Data Fig. 2), which is suggestive of deterministic Keplerian motion rather than the more stochastic stellar-activity variations.

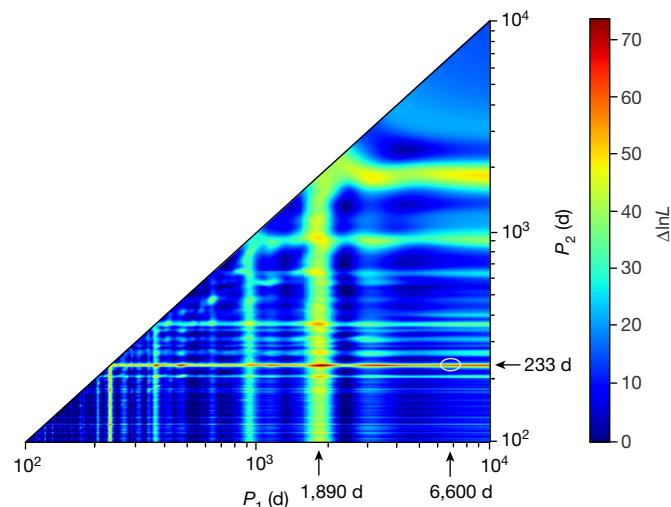


Fig. 1 | Two-dimensional likelihood periodogram. We used a multi-dimensional generalized Lomb–Scargle scheme assuming a white-noise model to explore combinations of periods to fit the data. The colour scale shows the improvement in the logarithm of the likelihood function $\Delta \ln L$ as a function of trial periods P_1 and P_2 . $\Delta \ln L > 18.1$ corresponds to a significant detection (FAP < 0.1%) for one signal, whereas two signals require $\Delta \ln L > 36.2$. The highest likelihood value ($\Delta \ln L = 71$) corresponds to periods of 233 days and 1,890 days, but all combinations of 233 days and periods longer than 2,500 days yield $\Delta \ln L > 65$ and are therefore statistically equivalent. The proposed solution discussed in the text ($P_1 = 233$ d and $P_2 = 6,600$ d) is indicated by a white ellipse.

Although stellar activity does not appear to be responsible for the 233-day signal in radial velocity, it could affect the significance and determination of the model parameters. We therefore carried out a study considering different models for correlated noise, based on moving averages and Gaussian processes. The moving-average models yield results that are comparable with the analysis assuming white noise and confirm the high statistical significance of the 233-day periodicity, with a FAP of 5×10^{-10} . The Gaussian-process framework strongly reduces the significance of the signal, with a FAP of no more than about 10%. However, Gaussian-process models have been shown¹⁴ to underestimate the significance of signals, even in the absence of correlated noise.

Despite the degeneracies encountered with certain models and after extensive testing (see Methods for further details), we conclude that the 233-day signal in the radial velocities is best explained as arising from a planet with minimum mass of 3.2 Earth masses in a low-eccentricity orbit with a semi-major axis of 0.40 AU. The median parameter values from our analysis are provided in Table 1 and Extended Data Table 2; Fig. 3 shows the models of the radial velocities. Standard Markov chain Monte Carlo (MCMC) procedures were used to sample the posterior distribution. The MCMC analysis yields a secular trend that is significantly different from zero. Both the trend and the long-term modulation could be related to a stellar-activity cycle (as photometric and spectroscopic indicators may suggest), but the presence of an outer planet cannot be ruled out. In the latter case, the fit suggests an object of more than about 15 Earth masses in an orbit with a semi-major axis of about 4 AU. This orbital period is compatible with that claimed previously⁶ from an astrometric long-term study, but the Doppler amplitude is inconsistent, unless the orbit is nearly face-on. On the other hand, the induced nonlinear astrometric signature over roughly 5 yr would be up to 3 mas, making it potentially detectable with the Gaia mission.

Extended Data Fig. 1 shows that some marginally significant signals might be present in the residuals of the two-signal model (for example, at 81 days), but current evidence is inconclusive. We can, however, set stringent limits on the exoplanet detectability in close-in orbits around Barnard's star. Our analysis is sensitive to planets with minimum masses of 0.7 and 1.2 Earth masses for orbital periods of 10 and 40 days, respectively, which correspond to the inner and outer

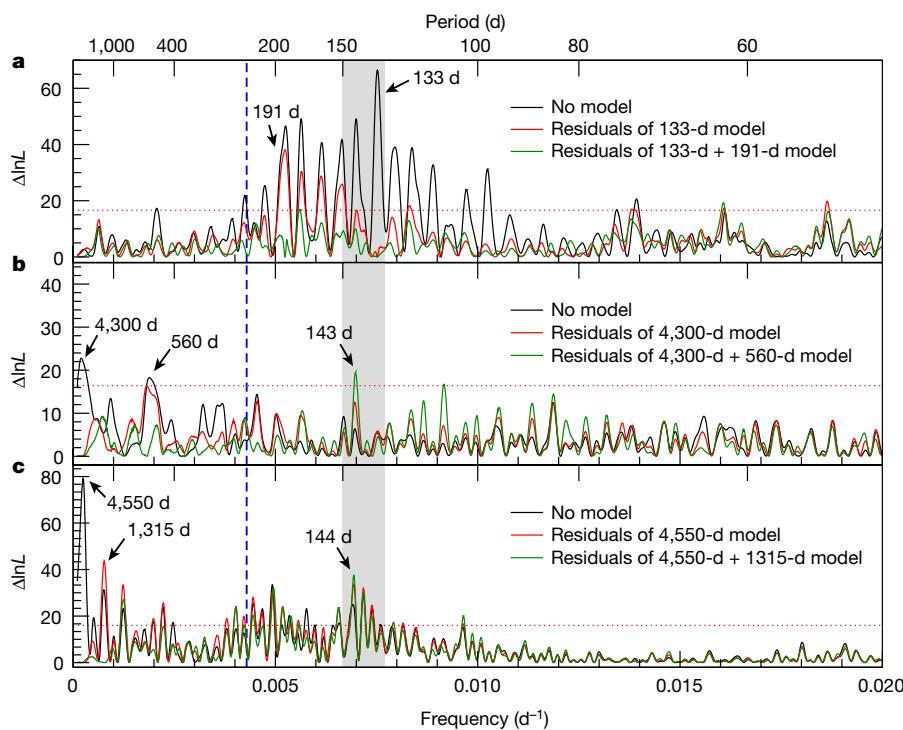


Fig. 2 | Periodicities in stellar activity indicators. a–c, Likelihood periodograms of time series in the central flux of the H α line (a), the emission in the Ca II H + K lines (b) and photometry (c). These indicators are associated with the presence of active regions on the stellar surface. Likelihood periodograms were obtained by including one signal at a time (sinusoids), as in the analysis of the radial velocities. The vertical dashed blue line indicates the location of the planetary signal from the radial-velocity analysis, at a period of 233 days; the dotted red line shows the detection threshold of FAP = 0.1%. The shaded region marks the most likely stellar rotation interval.

optimistic habitable-zone limits¹⁵. Barnard's star seems to be devoid of Earth-mass and larger planets and in hot and temperate orbits, in contrast with the seemingly high occurrence of planets in close-in orbits around M-type stars found by the Kepler mission^{16,17}.

The proximity of Barnard's star and the relatively large orbital separation makes the system ideal for astrometric detection. The Gaia and Hubble Space Telescope missions can reach an astrometric accuracy of 0.03 mas^{18,19}. Depending on the orbital inclination, they could detect the planetary signal or set a constraining upper limit on its mass (L.T.-O. et al., submitted manuscript). Also, for the calculated

orbital separation, the contrast ratio between the planet and the star in reflected light is of the order of 10^{-9} , depending on the adopted values of the geometric albedo and orbital inclination. This contrast ratio is beyond the capabilities of current imaging instrumentation by three orders of magnitude. However, the maximum apparent separation of 220 mas should be within reach of direct-imaging instruments planned for the next decade²⁰, which could potentially reveal a wealth of information.

The candidate planet lies almost exactly at the expected position of the snow line of the system, which is located at about 0.4 AU²¹. It has

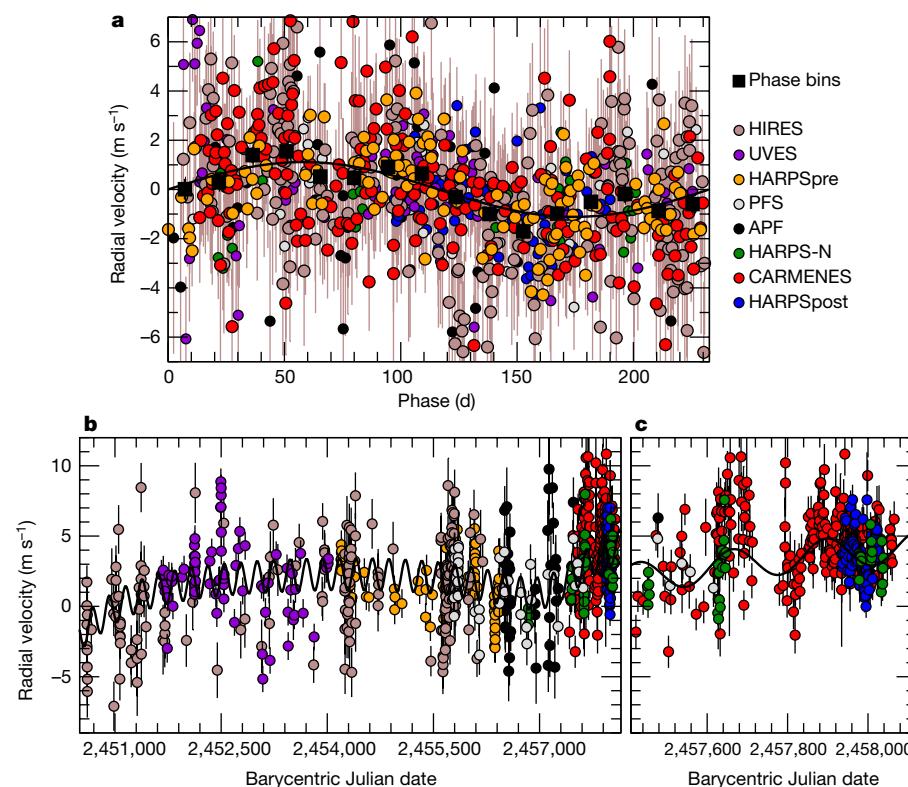


Fig. 3 | Fits to the radial-velocity time series. a, Phase-folded representation of the best-fitting 233-day circular orbit (black line) to the different datasets (circles; see Methods for dataset information and acronyms). The black squares represent the average velocity in 16 bins along the orbital phase. b, c, Time series of the radial-velocity observations with the fitted model superimposed (b) and a close-up of the time region around the CARMENES observations (c). The model fit (black line) corresponds to a solution assuming two signals (one of them forced to a period of more than 4,000 days, for reasons discussed in the text). In all panels, 1σ error bars on the measurements are shown (brown in b; black in b and c).

long been suggested that this region might provide a favourable location for forming planets^{22,23}, with super-Earths being the most common types of planet formed around low-mass stars²⁴. Recent models that incorporate dust coagulation, radial drift and planetesimal formation via the streaming instability support this idea²⁵. Although it has not yet been shown to be part of a general trend, observational evidence would substantially constrain theories of planetary migration²⁶.

The long-term intensive monitoring of Barnard's star and the precision of the measurements, which incorporate data from all precise, high-resolution spectrometers in operation, pushes the limits of the radial-velocity technique into a new regime of parameter space, namely super-Earth-type planets in cool orbits. This provides a bridge with the microlensing technique, which has traditionally been the only probe for small planets in orbits close to the snow line^{27,28}.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0677-y>.

Received: 16 March 2018; Accepted: 1 October 2018;

Published online 14 November 2018.

1. Brown, A. G. A. et al. Gaia data release 2: summary of the contents and survey properties. *Astron. Astrophys.* **616**, A1 (2018).
2. Liefke, C. & Schmitt, J. H. M. M. The NEXXUS database – X-ray properties of nearby stars. *ESA Spec. Publ.* **560**, 755–756 (2005).
3. Suárez Mascareño, A., Rebolo, R., González Hernández, J. I. & Esposito, M. Rotation periods of late-type dwarf stars from time series high-resolution spectroscopy of chromospheric indicators. *Mon. Not. R. Astron. Soc.* **452**, 2745–2756 (2015).
4. Zechmeister, M., Kürster, M. & Endl, M. The M dwarf planet search programme at the ESO VLT + UVES. A search for terrestrial planets in the habitable zone of M dwarfs. *Astron. Astrophys.* **505**, 859–871 (2009).
5. Choi, J. et al. Precise Doppler monitoring of Barnard's star. *Astrophys. J.* **764**, 131 (2013).
6. Bonfils, X. et al. The HARPS search for southern extra-solar planets. XXXI. The M-dwarf sample. *Astron. Astrophys.* **549**, A109 (2013).
7. van de Kamp, P. The planetary system of Barnard's star. *Vistas Astron.* **26**, 141–157 (1982).
8. Benedict, G. F. et al. Interferometric astrometry of Proxima Centauri and Barnard's star using Hubble Space Telescope Fine Guidance Sensor 3: detection limits for substellar companions. *Astron. J.* **118**, 1086–1100 (1999).
9. Gauza, B. et al. Constraints on the substellar companions in wide orbits around the Barnard's star from CanariCam mid-infrared imaging. *Mon. Not. R. Astron. Soc.* **452**, 1677–1683 (2015).
10. Anglada-Escudé, G. et al. A terrestrial planet candidate in a temperate orbit around Proxima Centauri. *Nature* **536**, 437–440 (2016).
11. Quirrenbach, A. et al. CARMENES instrument overview. *Proc. SPIE* **9147**, 91471F (2014).
12. Meschiari, S. et al. Systemic: a testbed for characterizing the detection of extrasolar planets. I. The systemic console package. *Publ. Astron. Soc. Pacif.* **121**, 1016–1027 (2009).
13. Benedict, G. F. et al. Photometry of Proxima Centauri and Barnard's star using Hubble Space Telescope Fine Guidance Sensor 3: a search for periodic variations. *Astron. J.* **116**, 429–439 (1998).
14. Feng, F., Tuomi, M., Jones, H. R. A., Butler, R. P. & Vogt, S. A Goldilocks principle for modelling radial velocity noise. *Mon. Not. R. Astron. Soc.* **461**, 2440–2452 (2016).
15. Kopparapu, R. K. et al. Habitable zones around main-sequence stars: dependence on planetary mass. *Astrophys. J.* **787**, L29 (2014).
16. Gaidos, E., Mann, A. W., Kraus, A. L. & Ireland, M. They are small worlds after all: revised properties of Kepler M dwarf stars and their planets. *Mon. Not. R. Astron. Soc.* **457**, 2877–2899 (2016).
17. Dressing, C. D. & Charbonneau, D. The occurrence of potentially habitable planets orbiting M dwarfs estimated from the full Kepler dataset and an empirical measurement of the detection sensitivity. *Astrophys. J.* **807**, 45 (2015).
18. Perryman, M., Hartman, J., Bakos, G. Á. & Lindegren, L. Astrometric exoplanet detection with Gaia. *Astrophys. J.* **797**, 14 (2014).
19. Casertano, S. et al. Parallax of Galactic Cepheids from spatially scanning the Wide Field Camera 3 on the Hubble Space Telescope: the case of SS Canis Majoris. *Astrophys. J.* **825**, 11 (2016).
20. Trauger, J. et al. Hybrid Lyot coronagraph for WFIRST-AFTA: coronagraph design and performance metrics. *J. Astron. Telesc. Instrum. Syst.* **2**, 011013 (2016).
21. Kennedy, G. M. & Kenyon, S. J. Planet formation around stars of various masses: the snow line and the frequency of giant planets. *Astrophys. J.* **673**, 502–512 (2008).

22. Stevenson, D. J. & Lunine, J. I. Rapid formation of Jupiter by diffuse redistribution of water vapor in the solar nebula. *Icarus* **75**, 146–155 (1988).
23. Morbidelli, A., Lambrechts, M., Jacobson, S. & Bitsch, B. The great dichotomy of the Solar System: small terrestrial embryos and massive giant planet cores. *Icarus* **258**, 418–429 (2015).
24. Mulders, G. D., Pascucci, I. & Apai, D. An increase in the mass of planetary systems around lower-mass stars. *Astrophys. J.* **814**, 130 (2015).
25. Drążkowska, J. & Alibert, Y. Planetesimal formation starts at the snow line. *Astron. Astrophys.* **608**, A92 (2017).
26. Kley, W. & Nelson, R. P. Planet-disk interaction and orbital evolution. *Annu. Rev. Astron. Astrophys.* **50**, 211–249 (2012).
27. Gaudi, B. S. Microlensing surveys for exoplanets. *Annu. Rev. Astron. Astrophys.* **50**, 411–453 (2012).
28. Suzuki, D. et al. The exoplanet mass-ratio function from the MOA-II survey: discovery of a break and likely peak at a Neptune mass. *Astrophys. J.* **833**, 145 (2016).
29. Pasquetti, V. M. et al. The CARMENES search for exoplanets around M dwarfs. Photospheric parameters of target stars from high-resolution spectroscopy. *Astron. Astrophys.* **615**, A6 (2018).

Acknowledgements The results are based on observations made with the CARMENES instrument at the 3.5-m telescope of the Centro Astronómico Hispano-Alemán de Calar Alto (CAHA, Almería, Spain), funded by the German Max-Planck-Gesellschaft (MPG), the Spanish Consejo Superior de Investigaciones Científicas (CSIC), the European Union and the CARMENES Consortium members; the 90-cm telescope at the Sierra Nevada Observatory (Granada, Spain) and the 40-cm robotic telescope at the SPACEOBOS observatory (San Pedro de Atacama, Chile), both operated by the Instituto de Astrofísica de Andalucía (IAA); and the 80-cm Joan Oró Telescope (TJO) of the Montsec Astronomical Observatory (OAdM), owned by the Generalitat de Catalunya and operated by the Institute of Space Studies of Catalonia (IEEC). This research was supported by the following institutions, grants and fellowships: Spanish MINECO ESP2016-80435-C2-1-R, ESP2016-80435-C2-2-R, AYA2016-79425-C3-1-P, AYA2016-79245-C3-2-P, AYA2016-79425-C3-3-P, AYA2015-69350-C3-2-P, ESP2014-54362-P, AYA2014-56359-P, RYC-2013-14875; Generalitat de Catalunya/CERCA programme; Fondo Europeo de Desarrollo Regional (FEDER); German Science Foundation (DFG) Research Unit FOR2544, project JE 701/3-1; STFC Consolidated Grants ST/P000584/1, ST/P000592/1, ST/M001008/1; NSF AST-0307493; Queen Mary University of London Scholarship; Perren foundation grant; CONICYT-FONDECYT 1161218, 3180405; Swiss National Science Foundation (SNSF); Koshland Foundation and McDonald-Leapman grant; and NASA Hubble Fellowship grant HST-HF2-51399.001. J.T. is a Hubble Fellow.

Reviewer information *Nature* thanks I. Snellen and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions I.R. led the CARMENES team and the TJO photometry, organized the analysis of the data and wrote most of the manuscript. M.T. performed the initial radial-velocity analysis and, with J.C.M., M.P., S.D., A.Ro., F.F., T.T., S.S.V., A.H., A.K., S.S.V., J.J. and A.S.M., participated in the analysis of radial-velocity data using various methods. A.Re. co-led the CARMENES team. R.P.B. led the HIRES/PFS/APF team and reprocessed the UVES data. C.R.-L. coordinated the acquisition and analysis of photometry. J.I.G.H., R.R., A.S.M. and B.T.-P. acquired HARPS-N data and measured chromospheric indices from all spectroscopic datasets. T.T. and M.H.L. studied the dynamics. S.S.V. co-led the HIRES/APF teams. J.A.C. is responsible for the CARMENES instrument and, with A.S. and M.C.-C., determined the stellar properties. E.H., F.M., E.R., J.B.P.S., S.G.E., E.F.G., M.Ki. and M.J.L.-G. participated in the photometric monitoring. S.V.J. contributed to the analysis of activity and to the preparation of the manuscript. M.L. calculated the cross-correlation function parameters of CARMENES spectra. R.N. participated in the discussion of implications for planet formation. A.Q. and P.J.A. are principal investigators of CARMENES. M.A., V.J.S.B., T.H., M.Ku., D.M., E.P., S.R. and W.S. are members of the CARMENES Consortium. L.T.-O. calibrated the CARMENES data and carried out calculations of astrometric detection. M.Z. reduced the CARMENES data. J.T., J.B., J.D.C., B.H., S.A.S. and S.X.W. participated in the acquisition and discussion of HIRES, PFS and APF data. J.R.B., G.C., C.A.H., J.J., H.R.A.J., J.M., A.O., D.S., R.A.S. and Y.T. participated in the RedDots2017 Collaboration. Z.M.B. participated in the discussion of instrument systematics. G.A.-E. organized the collaboration, coordinated the RedDots2017 campaign, organized and performed analyses and participated in writing the manuscript. All authors were given the opportunity to review the results and comment on the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0677-y>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0677-y>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to I.R.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Description of the individual radial-velocity datasets. As in other recent low-amplitude exoplanet discoveries, combining information from several instruments (historical data and quasi-simultaneous monitoring) is central to unambiguously identifying significant periodicities in the data. The suite of instruments used for this study and relevant information on the observation time intervals, the number of epochs, and the references of the observational programmes involved are provided in Extended Data Table 1.

The HIRES, PFS and APF datasets were obtained with the HIRES spectrometer³⁰ on the Keck I 10-m telescope atop Mauna Kea in Hawaii, the Planet Finding Spectrometer (PFS)³¹ on Carnegie's Magellan II 6.5-m telescope and the Automated Planet Finder (APF)³² on the 2.4-m telescope atop Mt Hamilton at Lick Observatory, respectively. In all cases, radial velocities were calibrated by placing a cell of gaseous iodine in the converging beam of the telescope, just ahead of the spectrometer slit. The iodine superimposes a rich forest of absorption lines on the stellar spectrum over the 5,000–6,200-Å region, thereby providing a wavelength calibration and proxy for the point spread function of the spectrometer. Once extracted, the iodine region of each spectrum is divided into 2-Å-wide chunks, resulting in about 700 chunks for APF and HIRES, and about 800 for PFS. Each chunk produces an independent measure of the absolute wavelength, point spread function and Doppler shift, determined using a previously described³³ spectral synthesis technique. The final reported Doppler velocity of each stellar spectrum is the weighted mean of the velocities of all the individual chunks. The final uncertainty of each velocity is the standard deviation of all chunk velocities about the weighted mean.

Further radial-velocity measurements of Barnard's star were obtained with the two HARPS spectrometers, ESO/HARPS³⁴ at the 3.6-m ESO telescope at La Silla Observatory and HARPS-N³⁵ at the 3.5-m Telescopio Nazionale Galileo in La Palma. These are high-resolution echelle spectrometers optimized for precision radial velocities covering a wavelength range of 3,800–6,800 Å. High stability is achieved by keeping the instrument thermally and mechanically isolated from the environment. All observations were wavelength-calibrated with emission lines of a hollow-cathode lamp and reduced using the pipeline Data Reduction Software. For the ESO/HARPS instrument, two distinct datasets are considered (HARPSpre, HARPSpost), corresponding to data acquired before and after a fibre upgrade that took place in June 2015. Radial velocities were obtained using the TERRA³⁶ software, which builds a high signal-to-noise template by co-adding all the existing observations and then performs a maximum likelihood fit of each observed spectrum against the template, yielding a measure of the Doppler shift and its uncertainty. The analysis of the initial HARPSpre dataset, which spans about six years, revealed a very prominent signal at a period compatible with one year. Thorough investigation led to the conclusion that this is a spurious periodicity caused by the displacement of the stellar spectrum on the detector over the year and the existence of physical discontinuities in the detector structure³⁷. We calculated new velocities by removing an interval of ± 45 km s⁻¹ around the detector discontinuities to account for the amplitude of Earth's barycentric motion. After this correction, all search analyses showed the one-year periodic signal disappearing well below the significance threshold, although some periodicity remains (possibly related to residual systematic effects in all datasets).

We also use radial-velocity measurements of Barnard's star obtained with the UVES spectrograph on the 8.2-m VLT UT2 at Paranal Observatory in the context of the M-dwarf programme executed between 2000 and 2008⁴. New radial-velocity measurements were obtained by reprocessing the iodine-based observations using up-to-date reduction codes¹⁰, as used in the HIRES, PFS and APF spectrometers.

Barnard's star was observed almost daily in the context of the CARMENES survey of rocky planets around red dwarfs³⁸, which uses the CARMENES instrument, a stabilized visible and near-infrared spectrometer on the 3.5-m telescope of Calar Alto Observatory. The data were pipeline-processed and radial velocities and their uncertainties were measured with the SERVAL algorithm³⁹, which is based on a template-matching scheme. For this study we used visual-channel radial velocities, which correspond to a wavelength interval of 5,200–9,600 Å. Because of instrument effects, data were further corrected by calculating a night-to-night offset (generally below 3 m s⁻¹) and a nightly slope (less than 3 m s⁻¹ peak to peak) from a large sample of observed stars. Barnard's star was excluded from the calibration to avoid biasing the results. The origin of the offsets is still unclear, but they are probably related to systematics in the wavelength solution, light scrambling and a slow drift in the calibration source during the night. After the corrections, CARMENES data have similar precision and accuracy to those from ESO/HARPS⁴⁰.

Barycentric correction, secular acceleration and other geometric effects. Although stellar motions on the celestial sphere are generally small, the measurement of precision radial velocities must carefully account for some perspective effects, including the motion of the target star and of the observer. This includes, in particular, secular acceleration⁴. A thorough description of a complete barycentric correction scheme down to a precision of less than 1 cm s⁻¹ is given elsewhere⁴¹.

We ensured that the barycentric corrections used in all our datasets agree with the code in ref. ⁴¹. Given its proximity to the Sun and high proper motion, Barnard's star is particularly susceptible to errors due to unaccounted terms in its motion. We systematically revised the apparent Doppler shifts accounting for the small but important changes in the apparent position over time.

Uncertainties in the astrometry (parallax, radial velocity and proper motion) could propagate into small residual signals in the barycentric correction. We performed numerical experiments to assess the effect of such uncertainties. Extended Data Fig. 3 shows the spurious one-year signal expected by introducing a shift of 150 mas (10 times larger than the uncertainties in the Hipparcos catalogue) in right ascension (RA) and declination (dec.) over a time-interval between years 2000 and 2018. The peak-to-peak amplitudes for such errors are roughly 4 cm s⁻¹. The next-largest terms are those that couple the proper motion with the tangential velocity of the star and of the observer. For this experiment we introduced errors of 15 mas yr⁻¹ in both proper motions in the direction of increasing RA and dec., and of 15 mas in the parallax (10 times larger than the uncertainties in the Hipparcos catalogue). The spurious signals caused by proper motion contain a trend (change in secular acceleration) and signal with a period of 1 yr growing in amplitude with time. The 1-yr periodicities are small and not significant, but the secular trend can produce detectable effects, mostly owing to the error in the parallax. The effect of errors at the 1 σ , 3 σ and 10 σ levels of Hipparcos uncertainties are shown in the bottom panel of Extended Data Fig. 3. Crucially, this signal consists of a trend that is easily included in the model without any major effect on the significance of the signal corresponding to the candidate planet.

Models and statistical tools. *Doppler model.* The Doppler measurements are modelled using the following equations:

$$v(t_i) = \gamma_{\text{INS}} + S(t_i - t_0) + \sum_{p=1}^n f_p(t_i) \\ f_p(t_i) = K_p \cos[\nu_p(t_i; P_p, M_{0,p}, e_p) + \varpi_p] + e_p \cos \varpi_p$$

where γ_{INS} (constant offset of each instrument) and S (linear trend) are free parameters. All signals are included in the Keplerian f_p , and for each planet K_p is the Doppler semi-amplitude, P_p is the orbital period, $M_{0,p}$ is the mean anomaly at t_0 , e_p is the orbital eccentricity and ϖ_p is the argument of periastron of the orbit. Precise definitions of the parameters and the calculation of the true anomaly ν_p can be found elsewhere⁴². In some cases, the orbits are assumed to be circular and the Keplerian term simplifies to

$$f_{p,\text{circ}}(t_i) = K_p \cos \left[\frac{2\pi}{P_p} (t_i - t_0) + M_{0,p} \right]$$

which has only three free parameters (K_p , P_p and $M_{0,p}$). This model is used in initial exploratory searches or when analysing time series that do not necessarily contain Keplerian signals (for example, activity proxies).

Statistical figure-of-merit. The fits to the data are obtained by finding the set of parameters that maximize the likelihood function L , which is the probability distribution of the data fitting the model. L can take slightly different forms depending on the noise model adopted. For measurements with normally distributed noise it can be written as

$$L = \frac{1}{(2\pi)^{N_{\text{obs}}/2}} |C|^{-1/2} \exp \left(-\frac{1}{2} \sum_{i=1}^{N_{\text{obs}}} \sum_{j=1}^{N_{\text{obs}}} r_i C_{ij}^{-1} r_j \right)$$

where $r_i = v_{i,\text{obs}} - v(t_i)$ is the residual of observation i , C_{ij} are the components of the covariance matrix between measurements, $|C|$ is its determinant and N_{obs} is the number of observations. Starting from this definition, there are three types of model that we consider.

White-noise model. If all observations are statistically independent from each other, all variability is included in $v(t_i)$ and the covariance matrix is diagonal. In this case, the logarithm of L simplifies to

$$\ln(L_W) = -\frac{N_{\text{obs}}}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{N_{\text{obs}}} \ln(\varepsilon_i^2 + s_{\text{INS}}^2) - \frac{1}{2} \sum_{i=1}^{N_{\text{obs}}} \frac{r_i^2}{\varepsilon_i^2 + s_{\text{INS}}^2}$$

where ε_i is the nominal uncertainty of each measurement and s_{INS} is an excess noise component (often called the jitter parameter) for each instrument. We call this model the white-noise model because it implicitly assumes that the noise has a uniform power distribution in frequency space.

Moving average. Auto-regressive moving-average (ARMA) models can also be used⁴³ when measurements depend on the previous ones in a way that is difficult to parameterize with deterministic functions (for example, quasi-periodic variability, Brownian motion and impulsive events). In our case, we use an ARMA model

containing only one moving-average term assuming that each measurement is related to the previous residual as

$$r_{i,MA} = v_{i,obs} - [v(t_i) + r_{i-1,MA} \alpha_{INS} e^{-(t_i - t_{i-1})/\tau_{INS}}]$$

This model contains two additional parameters for each instrument: the coefficient α_{INS} and the timescale τ_{INS} , which represent the strength and time-coherence of the correlated noise, respectively⁴⁴.

Gaussian process. Finally, the most general model, often called a Gaussian process, involves parameterizing the covariance matrix⁴⁵:

$$C_{ij}^2 = s_{INS}^2 \delta_{ij} + \kappa(\tau_{ij})$$

where κ is the kernel function, which is a function of the time difference between observations $\tau_{ij} = |t_i - t_j|$ and some other free parameters. Many kernel functions exist with different properties. Here we consider a stochastically driven, damped simple harmonic oscillator⁴⁶ (SHO):

$$\kappa(\tau) = C_0 e^{-\tau/P_{life}} \times \begin{cases} \cosh\left(\eta \frac{2\pi\tau}{P_{rot}}\right) + \frac{P_{rot}}{2\pi\eta P_{life}} \sinh\left(\eta \frac{2\pi\tau}{P_{rot}}\right) & \text{for } P_{rot} > 2\pi P_{life} \\ 2\left(1 + \frac{2\pi\tau}{P_{rot}}\right) & \text{for } P_{rot} = 2\pi P_{life} \\ \cos\left(\eta \frac{2\pi\tau}{P_{rot}}\right) + \frac{P_{rot}}{2\pi\eta P_{life}} \sin\left(\eta \frac{2\pi\tau}{P_{rot}}\right) & \text{for } P_{rot} < 2\pi P_{life} \end{cases}$$

where P_{rot} is the stellar rotation period, P_{life} is the lifetime of active regions, C_0 is a scaling factor proportional to the fraction of the stellar surface covered by active regions, and $\eta = |1 - (2\pi P_{life}/P_{rot})^{-2}|^{1/2}$. This model is popular in astrophysical applications because its three parameters can be associated to physical properties. *False-alarm probability.* We use the frequentist concept of false-alarm probability of detection (FAP) to assess statistical significance. FAP is formally equivalent to the P used in other applications. The statistical significance of the detection of a planet is a problem of null hypothesis significance test, where the null hypothesis is a model with n signals (null model), and the model to be benchmarked contains $n + 1$ signals with a correspondingly larger number of parameters. The procedure is as follows:

First, we compute $\ln L$ of the null model, which contains all n detected signals and nuisance parameters (jitters, trend, and so on).

Second, $\ln L$ is maximized by adjusting all the model parameters together with the parameters of a sinusoid for a list of test periods for signal $n + 1$. Then, the logarithm of the improvement in the likelihood function with respect to the null model is computed ($\Delta \ln L_{P_{n+1}}$) at each test period P and plotted against the values for all other periods to generate a log-likelihood periodogram⁴⁷.

Third, the largest $\Delta \ln L_{P_{n+1}}$ (the peak in the periodogram) indicates the most favoured period for the new signal. This value is then compared with the probability of randomly finding such an improvement when the null hypothesis is true, which is the desired FAP⁴⁸. A FAP around 1% would be considered tentative evidence, and below 10^{-3} (or 0.1%) is considered statistically significant.

All FAP assessments and significances presented here, including Doppler data and activity indicators, are computed using this procedure. We note that FAPs depend on the noise model that we adopt (white noise, moving average or Gaussian process).

Bayesian tools and analyses. We also applied Bayesian criteria to the detection of signals (Bayesian factors as in ref. ¹⁴), but these lead to conclusions and discussions qualitatively similar to those presented using frequentist criteria, so are omitted for brevity.

Median values and credibility intervals presented in tables were determined using a standard custom-made code implementing an MCMC algorithm⁴⁹. In all cases, uniform priors in all the parameters were assumed, with the exception of the periods. In that case, the prior was chosen to be uniform in frequency and an upper limit to the period was set to twice the timespan of the longest dataset (about 12,000 days).

Noise models and experiments applied to our datasets. If the presence of spurious Doppler variability caused by stellar activity is suspected, then checking the significance of the detections under different assumptions about the noise is advisable⁵⁰. The significance assessments in the main manuscript are given assuming a moving-average model for the radial-velocity analyses, and white-noise models for all other sets (photometry and activity indices). This section provides the justification for such an assumption. White-noise models are good for preliminary assessments but are prone to false positives¹⁴. On the other hand, Gaussian processes tend to produce overly conservative significance assessments leading to false negatives.

We investigated the performances of the different noise models by analysing the combination of three datasets in more detail: HIRES, HARPSpre and CARMENES. These are the relevant ones because they contribute most decisively to the improvement in the likelihood statistic (largest number of points, widest timespan and higher precision). The white-noise model found the signal at 233 days with $\Delta \ln L = 42$ ($\text{FAP} \approx 3.3 \times 10^{-14}$) and the moving-average model yielded a detection with $\Delta \ln L = 22.3$ ($\text{FAP} \approx 8.6 \times 10^{-6}$). On the other hand, a Gaussian process using the SHO kernel yielded a detection with only $\Delta \ln L = 11.6$ ($\text{FAP} \approx 27\%$). Despite this rather poor significance, Gaussian processes account for all covariances including those produced from real signals, which prompted us to carry out a deeper assessment.

We performed simulations by injecting a signal at 233 days (1.2 m s^{-1}) and attempted the detection using the three noise models. We first generated a synthetic sinusoidal signal (no eccentricity) and sampled it at the observing dates of the three sets. Random white-noise errors were then associated with each measurement in accordance with their formal uncertainties and the jitter estimates for each set. When using white-noise and moving-average models, a one-planet search trivially detected the signal at 233 days yielding $\Delta \ln L = 43$ ($\text{FAP} \approx 1.22 \times 10^{-14}$) and $\Delta \ln L = 32$ ($\text{FAP} \approx 6.3 \times 10^{-10}$), respectively, indicating high statistical significance. On the other hand, adding one planet when using Gaussian processes led to $\Delta \ln L = 14$ ($\text{FAP} \approx 2.7\%$), indicating that an unconstrained Gaussian process (all parameters free) absorbed $\Delta \ln L \approx 29$, even in the absence of any true correlated noise. This reduction is comparable to that observed in the real dataset (from $\Delta \ln L = 42$ for the white-noise model to $\Delta \ln L = 11.6$ when using a Gaussian-process model as discussed earlier), which supports the hypothesis that the Gaussian process is substantially absorbing the real signal, even if its parameters are set to match the rotation period of the star derived from spectroscopic indices and photometry (see Extended Data Fig. 4 for a visual representation of the effect).

The filtering properties of Gaussian processes can be better understood in Fourier space (frequency domain). As discussed previously⁴⁶, Gaussian processes fit for covariances within a range of frequencies filtered by the power spectral distribution (PSD) of the kernel function used. In particular, for an SHO kernel, the PSD is centred at the frequency of the oscillator, $\nu = 2\pi/P_{rot}$, and has a full-width at half-maximum of $2/P_{life}$. The activity indices of Barnard's star imply that ν and $2/P_{life}$ are comparable and of the order of 10^{-2} per day. Consequently, the Gaussian process strongly absorbs power (that is, $\Delta \ln L$) from signals in the frequency range $10^{-2} \pm 10^{-2} \text{ d}^{-1}$, which spans periods from 50 days to infinity, as illustrated by the black line in Extended Data Fig. 4. Most of the kernels proposed in the literature are very similar to the SHO kernel, so similar filtering properties are to be expected.

In a separate set of simulations, we checked the sensitivity of the three noise models to false positives by creating synthetic data from covariances. The results are in general agreement with previous results¹⁴ in the sense that the moving-average models have best statistical power. Furthermore, 300,000 datasets were generated using MCMC sampling of the SHO parameters. P_{rot} and P_{life} pairs were derived from MCMC fits to the $H\alpha$ time series and the corresponding C_0 parameters were obtained from an empirical relationship obtained from fitting Gaussian-process kernels with fixed P_{rot} and P_{life} to our real radial-velocity datasets. Next, synthetic observations were obtained using a multivariate random-number generator from the covariance matrix for all epochs. Reported uncertainties and jitter estimates for each observational dataset were added in quadrature and consistent white noise was also injected. Finally, a synthetic set was accepted only if it had a root-mean-square within 0.1 m s^{-1} of the real value. We then performed a maximum likelihood search using the moving-average model, and the solution with maximum likelihood was recorded in each case. This process produced a distribution of false alarms as a function of $\Delta \ln L$ and P_{rot} (Extended Data Fig. 5). This leads to $\text{FAP} \approx 0.8\%$ for our candidate signal. Although this is not an extremely low value, we consider it sufficiently small to claim a detection given that we followed a rather conservative procedure, and given the existing degeneracies between signals and correlated noise models. If we had carried out a deep scrutiny of each of the false alarms as we did with our real dataset, we would have discarded the fraction that failed our sanity checks (steady growth in signal strength, existence of a significant signal in populated dataset pairs, consistent offsets in overlapping regions, and so on). This would reduce the FAP estimated using this procedure.

In summary, we find that the most adequate models to account for the noise and to maximize the detection efficiency in this period domain are those that use moving-average terms, and that the 233-day signal is statistically significant under these models.

Zero-points between datasets. Calculating the zero points between the different datasets is key to ensure unbiased results and detection of genuine signals and to avoid introducing spurious effects. The best-fitting model is a self-consistent fit of the datasets allowing for a variable zero-point offset that is optimized via

maximum likelihood together with the search for periodicities. To validate these results, we used a complementary approach based on searching for overlapping coverage (typically a few nights) between different datasets to calculate average differences and thus measure zero-point offsets directly. We worked recursively, piecing datasets together one by one depending on the existence and size of overlap regions. We optimized the averaging window and selected the one that provided the best agreement in a three-way comparison. This is a trade-off between window size, number of points and measurement error. Periods below the window duration are affected by this process but our focus lies in a period of 233 days. Any window size smaller than a few tens of days does not affect the results.

The window parameters and the differences between the manually computed zero-point offsets and the values resulting from the optimization routine (considering a long-period signal) are given in Extended Data Table 3. The compatibility of the zero points calculated using two completely independent methods is very good. Only for UVES does a difference larger than 1σ appear. This can be attributed to the sparse sampling of the observations leading to small overlap between the datasets. Also, the zero point is based on a few measurements from HIRES that appear to deviate systematically from the average. Because of the reduced overlap, the resulting zero point is critically dependent on the window size and thus unreliable. The most populated datasets (HIRES, HARPSpre and CARMENES) have excellent zero-point consistency. In addition, the agreement of the general offsets of the combined set1 (HIRES, UVES, HARPSpre, APF and PFS) and set 2 (CARMENES, HARPS-N and HARPSpost) is remarkable (Extended Data Table 3). This is related to the presence of the long-term signal, which is found naturally when calculating manual offsets and confirmed from the global optimization including a long-period prior.

Stellar-activity analysis. Barnard's star is considered to be an aged, inactive star, but it appears to have small changing spots that make its rotation period tricky to ascertain. Spectroscopic indices ($H\alpha$ and $Ca\,\text{II}\,H + K$) and photometric measurements were used to estimate the period range in which signals from stellar activity are present. In all cases, the modelling of the data was performed using the same methodology as for the radial velocities, including the optimization of zero-point offsets and jitter terms for the different instruments, but assuming sinusoidal signals (zero eccentricity). As a result of the analysis, the stellar rotation period can be constrained to the range 130–150 d from all indicators, and there is also evidence for long-period modulation, which could be related to an activity cycle. No significant variability related to magnetic activity is present around 233 days, where the main radial-velocity periodic signal is found. A thorough review and analysis of all data on activity indicators for Barnard's star will be presented elsewhere.

Spectroscopy, $H\alpha$ index. Stellar activity was studied using the available spectroscopic data on Barnard's star. The $H\alpha$ index was calculated using three narrow spectral ranges covering the full $H\alpha$ line profile and two regions on the pseudo-continuum at both sides of the line, after normalizing the spectral order with a linear fit³. The error bars were estimated by adopting the standard deviation of the fluxes in a small local continuum region close the core of the lines as the uncertainty of the individual fluxes. The $H\alpha$ index was measured in 618 night-averaged spectra acquired with seven different instruments covering a timespan of 14.5 years. The analysis of the resulting time series (Fig. 2) yields a high-significance ($FAP \ll 0.1\%$) periodic signal at 133 days, and a second, also highly-significant signal at 191 days. We interpret the 133-day periodicity as tracing the stellar rotation period. This value is in relatively good agreement with a previous determination of 148 days³. The longer-period signal could be a consequence of the non-sinusoidal nature of the variability, the finite lifetime of active regions or the presence of differential rotation. The analysis of the $H\alpha$ index does not reveal any significant long-term (more than 1,000 days) modulation.

Spectroscopy, S-index. The S-index⁵¹ derived from the $Ca\,\text{II}\,H + K$ lines was only available for five instruments (APF, HARPS-N, HARPSpost, HARPSpre and HIRES). The S-index was estimated from 384 night-averaged spectra covering a similar timespan as for $H\alpha$. Two long-period signals were extracted from the analysis of the time series (Fig. 2), at periods of 4,300 days and 560 days. The next strongest signal, with $FAP \approx 10^{-4}$, has a period of 143 days and is probably associated with stellar rotation. Using an empirical relationship⁵², the activity-induced radial-velocity signal corresponding to this rotation period is predicted to be about 0.6 m s^{-1} . The long-term signal found from the S-index is consistent with estimates of activity cycles from photometric time series in other M stars of similar activity levels⁵³.

Photometry. Photometry from the literature includes data from the All Sky Automated Survey (ASAS)⁵⁴ and the MEarth Project⁵⁵. We also used unpublished photometry from the 0.8-m Four College Automated Photoelectric Telescope (FCAPT, Fairborn Observatory, Arizona, USA) and the 1.3-m Roboticly-Controlled Telescope (RCT, Kitt Peak National Observatory, Arizona, USA). In addition, new observations were acquired within the RedDots2017 campaign (<https://reddots.space/>) from the following facilities: the 0.90-m telescope at Sierra Nevada Observatory (Granada, Spain), the robotic 0.8-m Joan Oró telescope

(TJO, Montsec Astronomical Observatory, Lleida, Spain), Las Cumbres Observatory network with the 0.4-m telescopes located in Siding Spring Observatory, Teide Observatory and Haleakala Observatory, the ASH2 0.40-m robotic telescope at San Pedro de Atacama (Celestial Explorations Observatory, SPACEOBS, Chile), and from 14 observers of the American Association of Variable Stars Observers (AAVSO). A comprehensive summary of these measurements and contributors will be given elsewhere. The data cover about 15.1 years of observations with 1,634 epochs, a root-mean-square of 13.6 mmag and a mean error of 9.8 mmag. The analysis of the combined datasets (Fig. 2) indicates long-term modulations of 4,500 days and 1,300 days (semi-amplitudes of 10 mmag and 5 mmag, respectively) and two significant periods at 144 days and 201 days (semi-amplitudes of about 3 mmag). The interpretation is that the long-term modulation may be caused by an activity cycle whereas the signals at 144 days and 201 days are probably related to the base stellar rotation period and to the effects of the finite lifetime of active regions and differential rotation at different latitudes. The resulting periods are consistent with the results from the spectroscopic indices. A rotation period of 130.4 days and semi-amplitude of about 5 mmag had been reported previously¹³ from photometric observations, albeit with low significance ($FAP \approx 10\%$).

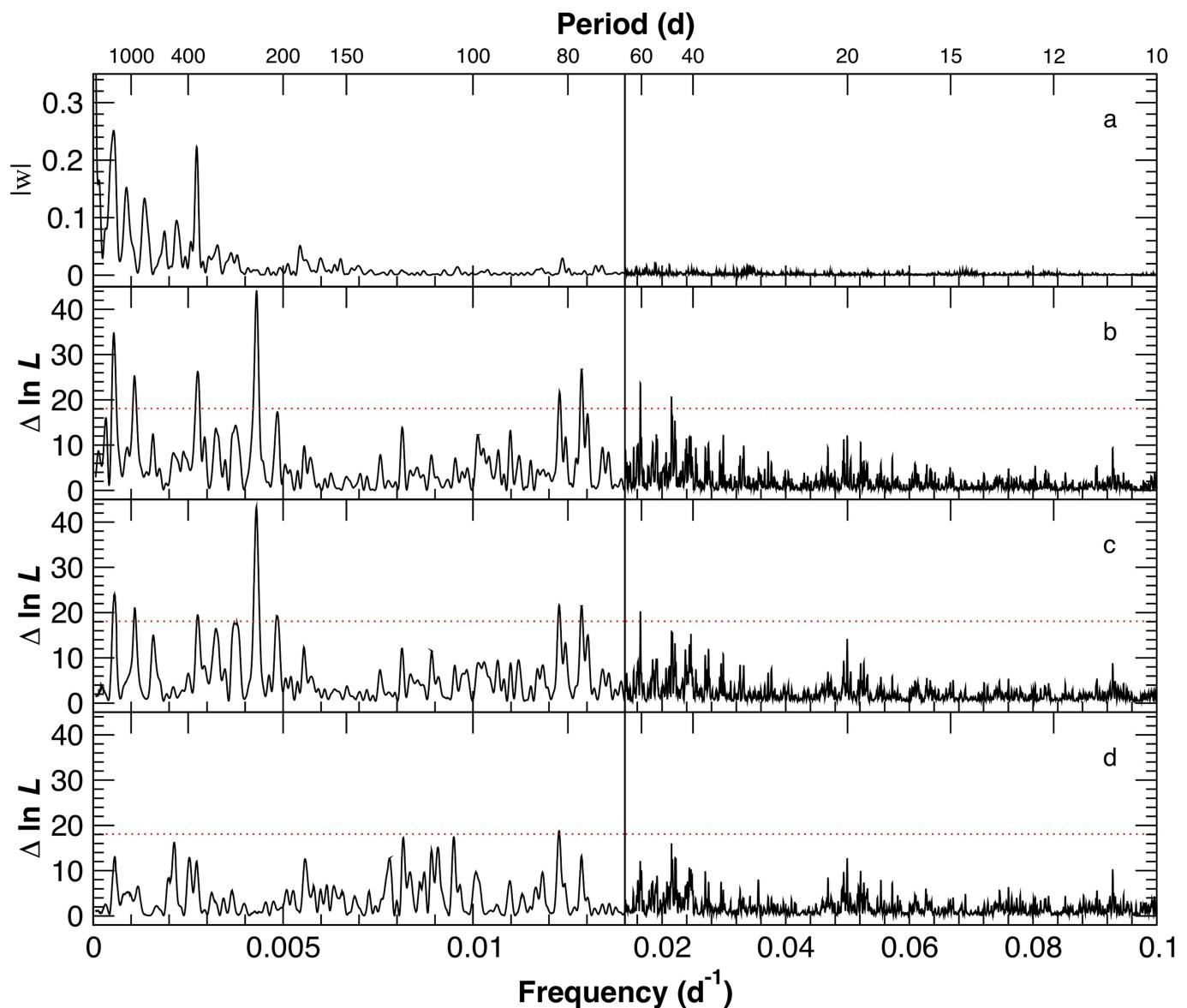
Code availability. The SERVAL template-matching radial-velocity measurement tool used for CARMENES data can be found at <https://github.com/mzechmeister/serval>. The TERRA template-matching radial-velocity measurement tool and various custom periodogram analysis and MCMC tools are codes written in Java by G.A.-E. and are available upon request (guillem.anglada@gmail.com). Other public codes and facilities used to model the data include GLS (<http://www.astro.physik.uni-goettingen.de/~zechmeister/gls.php>), Systemic Console (<https://github.com/stefano-meschieri/Systemic-Live>), Agatha (<https://github.com/phillippro/agatha>), Celerite (<https://github.com/dfm/celerite.git>) and EMCEE (<https://github.com/dfm/emcee>).

Data availability

The public high-resolution spectroscopic raw data used in the study can be freely downloaded from the corresponding facility archives: HIRES, <http://koa.ipac.caltech.edu>; UVES, HARPSpre and HARPSpost, <http://archive.eso.org>; HARPS-N, <http://archives.ia2.inaf.it/tng>; APF, <https://mthamilton.ucolick.org/data>). Proprietary raw data are available from the corresponding author on reasonable request. The nightly averaged, fully calibrated radial velocities, spectroscopic indices and photometric measurements are available as Supplementary Data.

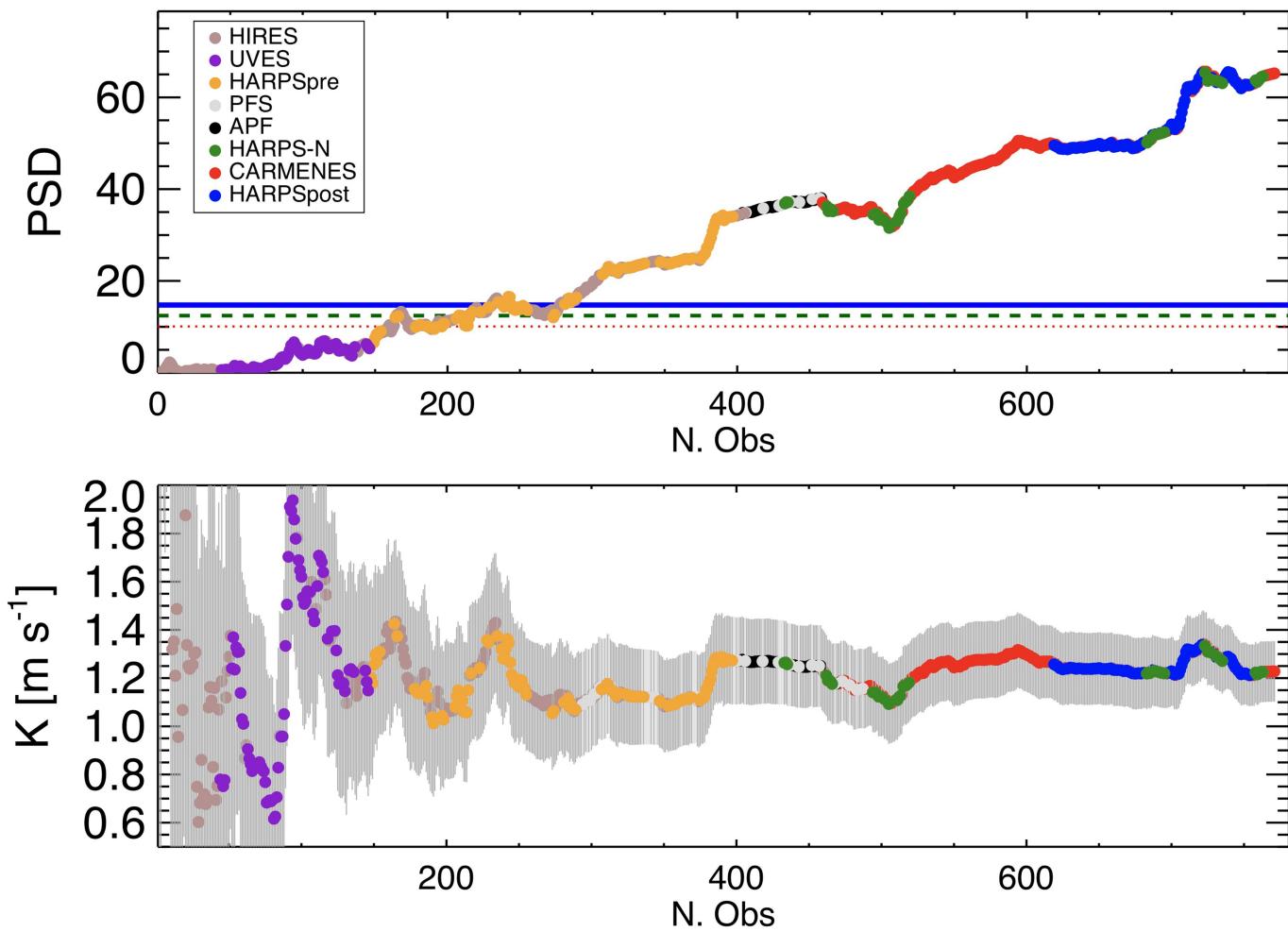
30. Vogt, S. S. et al. HIRES: the high-resolution Echelle spectrometer on the Keck 10-m telescope. *Proc. SPIE* **2198**, 362 (1994).
31. Crane, J. D. et al. The Carnegie planet finder spectrograph: integration and commissioning. *Proc. SPIE* **7735**, 773553 (2010).
32. Vogt, S. S. et al. APF – the Lick Observatory automated planet finder. *Publ. Astron. Soc. Pacif.* **126**, 359–379 (2014).
33. Butler, R. P. et al. Attaining Doppler precision of 3 m s^{-1} . *Publ. Astron. Soc. Pacif.* **108**, 500–509 (1996).
34. Mayor, M. et al. Setting new standards with HARPS. *Messenger* **114**, 20–24 (2003).
35. Cosentino, R. et al. HARPS-N: the new planet hunter at TNG. *Proc. SPIE* **8446**, 84461V (2012).
36. Anglada-Escudé, G. & Butler, R. P. The HARPS-TERRA project. I. Description of the algorithms, performance, and new measurements on a few remarkable stars observed by HARPS. *Astrophys. J. Suppl. Ser.* **200**, 15 (2012).
37. Dumusque, X., Pepe, F., Lovis, C. & Latham, D. W. Characterization of a spurious one-year signal in HARPS data. *Astrophys. J.* **808**, 171 (2015).
38. Quirrenbach, A. et al. CARMENES: an overview six months after first light. *Proc. SPIE* **9908**, 990812 (2016).
39. Zechmeister, M. et al. Spectrum radial velocity analyser (SERVAL). High-precision radial velocities and two alternative spectral indicators. *Astron. Astrophys.* **609**, A12 (2018).
40. Trifonov, T. et al. The CARMENES search for exoplanets around M dwarfs. First visual-channel radial-velocity measurements and orbital parameter updates of seven M-dwarf planetary systems. *Astron. Astrophys.* **609**, A117 (2018).
41. Wright, J. T. & Eastman, J. D. Barycentric corrections at 1 cm s^{-1} for precise Doppler velocities. *Publ. Astron. Soc. Pacif.* **126**, 838–852 (2014).
42. Lucy, L. B. Spectroscopic binaries with elliptical orbits. *Astron. Astrophys.* **439**, 663–670 (2005).
43. Scargle, J. D. Studies in astronomical time series analysis. I. Modeling random processes in the time domain. *Astrophys. J. Suppl. Ser.* **45**, 1–71 (1981).
44. Tuomi, M. et al. Habitable-zone super-Earth candidate in a six-planet system around the K2.5V star HD 40307. *Astron. Astrophys.* **549**, A48 (2013).
45. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, 2006).
46. Foreman-Mackey, D., Agol, E., Ambikasaran, S. & Angus, R. Fast and scalable Gaussian process modeling with applications to astronomical time series. *Astron. J.* **154**, 220 (2017).
47. Baluiev, R. V. Accounting for velocity jitter in planet search surveys. *Mon. Not. R. Astron. Soc.* **393**, 969–978 (2009).
48. Baluiev, R. V. Assessing the statistical significance of periodogram peaks. *Mon. Not. R. Astron. Soc.* **385**, 1279–1285 (2008).

49. Ford, E. B. Improving the efficiency of Markov chain Monte Carlo for analyzing the orbits of extrasolar planets. *Astrophys. J.* **642**, 505–522 (2006).
50. Dumusque, X. Radial velocity fitting challenge. I. Simulating the data set including realistic stellar radial-velocity signals. *Astron. Astrophys.* **593**, A5 (2016).
51. Duncan, D. K. et al. Ca ii H and K measurements made at Mount Wilson Observatory, 1966–1983. *Astrophys. J. Suppl. Ser.* **76**, 383–430 (1991).
52. Suárez Mascareño, A. et al. HADES RV programme with HARPS-N at TNG. VII. Rotation and activity of M-Dwarfs from time-series high-resolution spectroscopy of chromospheric indicators. *Astron. Astrophys.* **612**, A89 (2018).
53. Suárez Mascareño, A., Rebolo, R. & González Hernández, J. I. Magnetic cycles and rotation periods of late-type stars from photometric time series. *Astron. Astrophys.* **595**, A12 (2016).
54. Pojmański, G., Pilecki, B. & Szczygiet, D. The All Sky Automated Survey. Catalog of variable stars. V. Declinations 0° – $+28^\circ$ of the Northern Hemisphere. *Acta Astron.* **55**, 275–301 (2005).
55. Berta, Z. K., Irwin, J., Charbonneau, D., Burke, C. J. & Falco, E. E. Transit detection in the MEarth survey of nearby M dwarfs: bridging the clean-first, search-later divide. *Astron. J.* **144**, 145 (2012).
56. Zechmeister, M. & Kürster, M. The generalised Lomb-Scargle periodogram. A new formalism for the floating-mean and Keplerian periodograms. *Astron. Astrophys.* **496**, 577–584 (2009).
57. Affer, L. et al. HADES RV program with HARPS-N at the TNG GJ 3998: an early M-dwarf hosting a system of super-Earths. *Astron. Astrophys.* **593**, A117 (2016).
58. Mortier, A. & Collier Cameron, A. Stacked Bayesian general Lomb-Scargle periodogram: identifying stellar activity signals. *Astron. Astrophys.* **601**, A110 (2017).



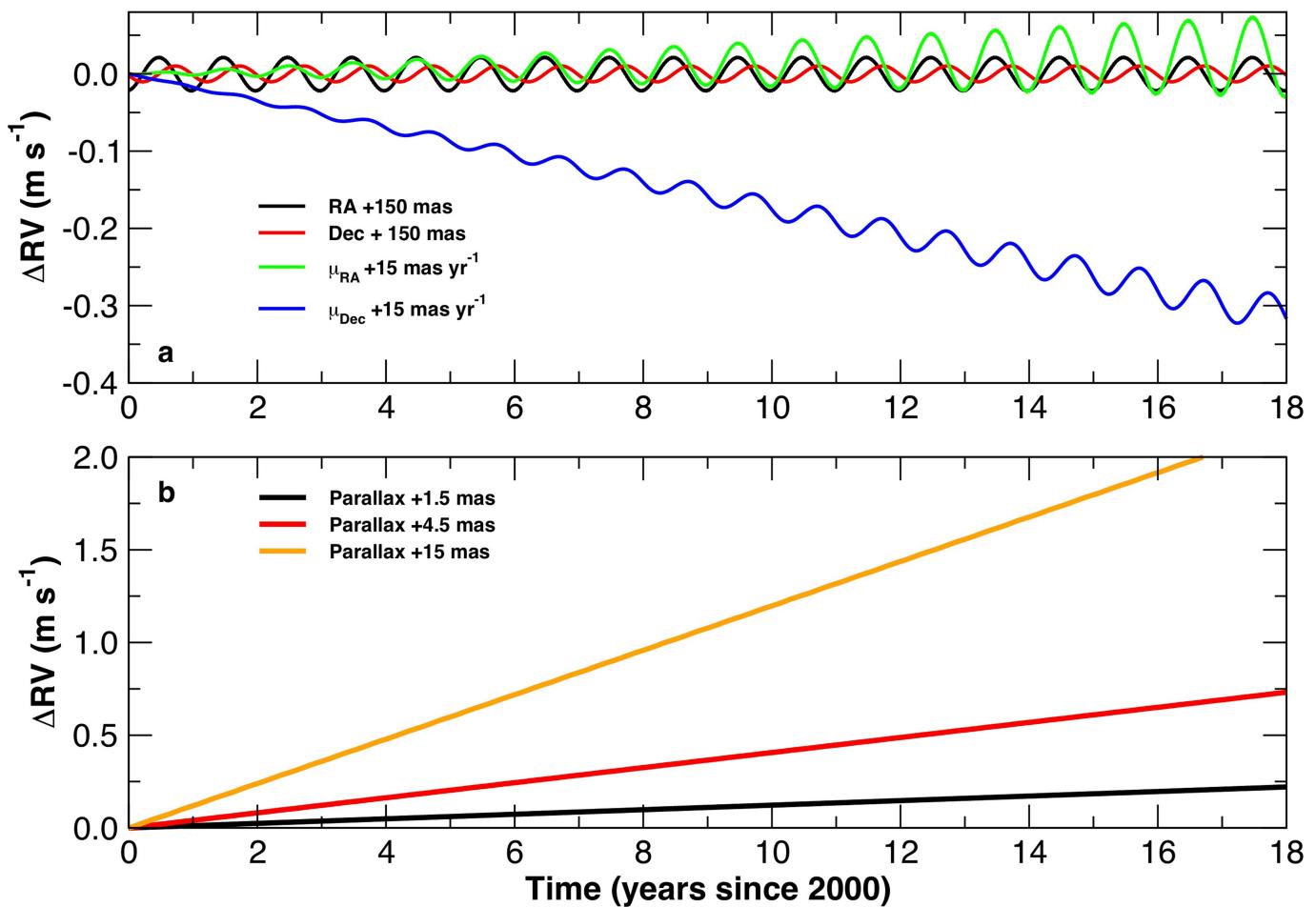
Extended Data Fig. 1 | Hierarchical periodogram analysis. **a**, Magnitude of the window function $|w|$ of the combined datasets. **b–d**, Likelihood periodogram of the radial-velocity measurements considering the first signal search (**b**), the residuals after modelling a long-period (6,600 days) signal (**c**) and the residuals after modelling long-period and 233-day periodicities (**d**). No high-significance signals remain in **d**, in particular in

the 10–40-day region, corresponding to the conservative habitable zone. The region below 10 days is not shown for clarity, but it is also devoid of significant periodic signals down to the Nyquist frequency of the dataset (2 days). Two different scales for the horizontal axis are used to improve the visibility of the low-frequency range. The red dotted line marks the 0.1% FAP threshold.



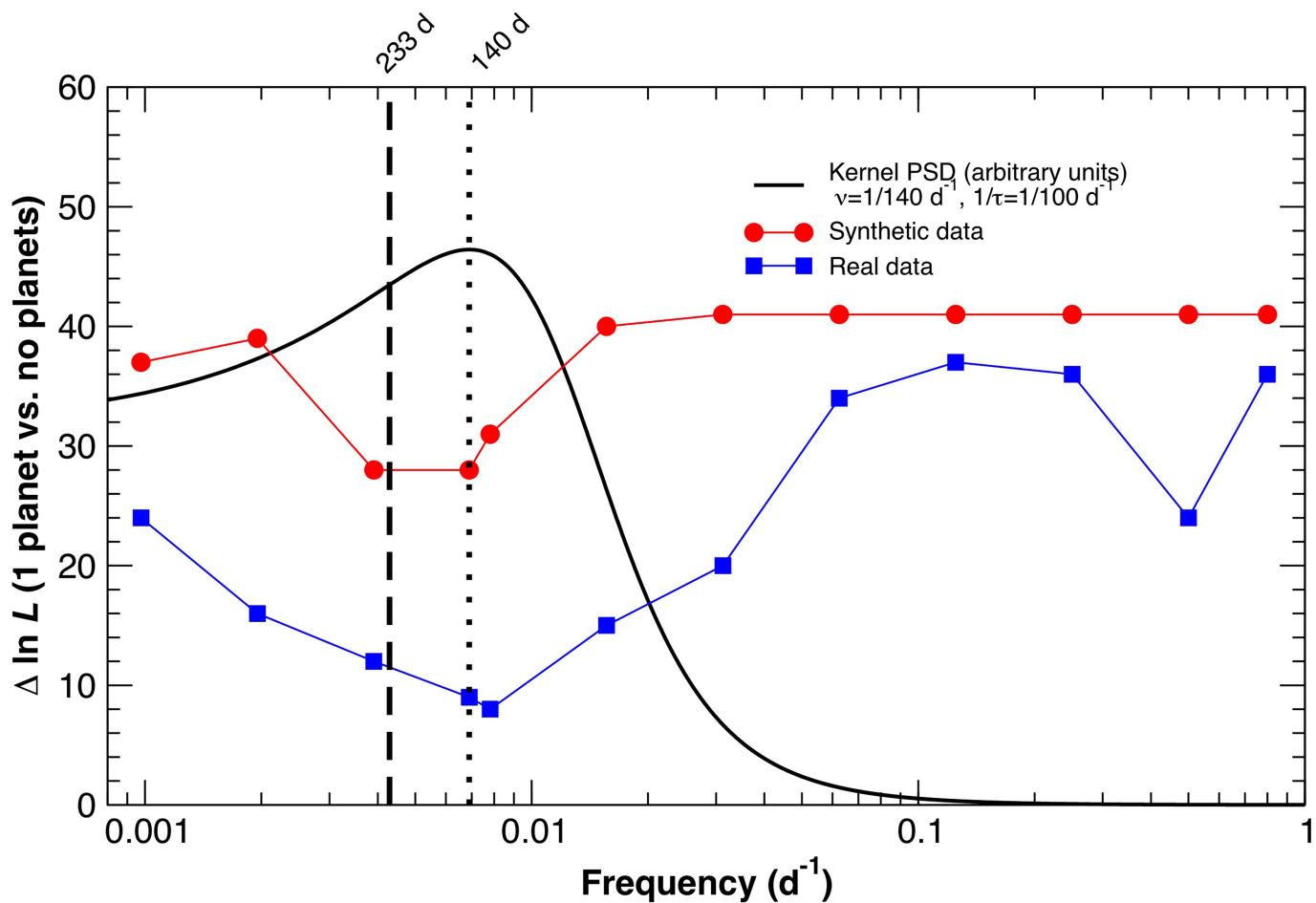
Extended Data Fig. 2 | Evolution of the significance of the 233-day signal. The top panel shows the PSD⁵⁶ of a stacked periodogram^{57,58} and the bottom panel depicts a cumulative measurement of the semi-amplitude K of the signal, with the grey lines showing 1σ error bars. The horizontal red dotted line, green dashed line and blue solid lines show the 10%, 1% and 0.1% FAP thresholds. The evolution of the significance is stable

with time and the variations in the amplitude over the last nine years of observations are smaller than 5% of the measured amplitude. The steady increase in signal significance and the stable amplitude are both consistent with the expected evolution of the evidence for a signal of Keplerian origin.



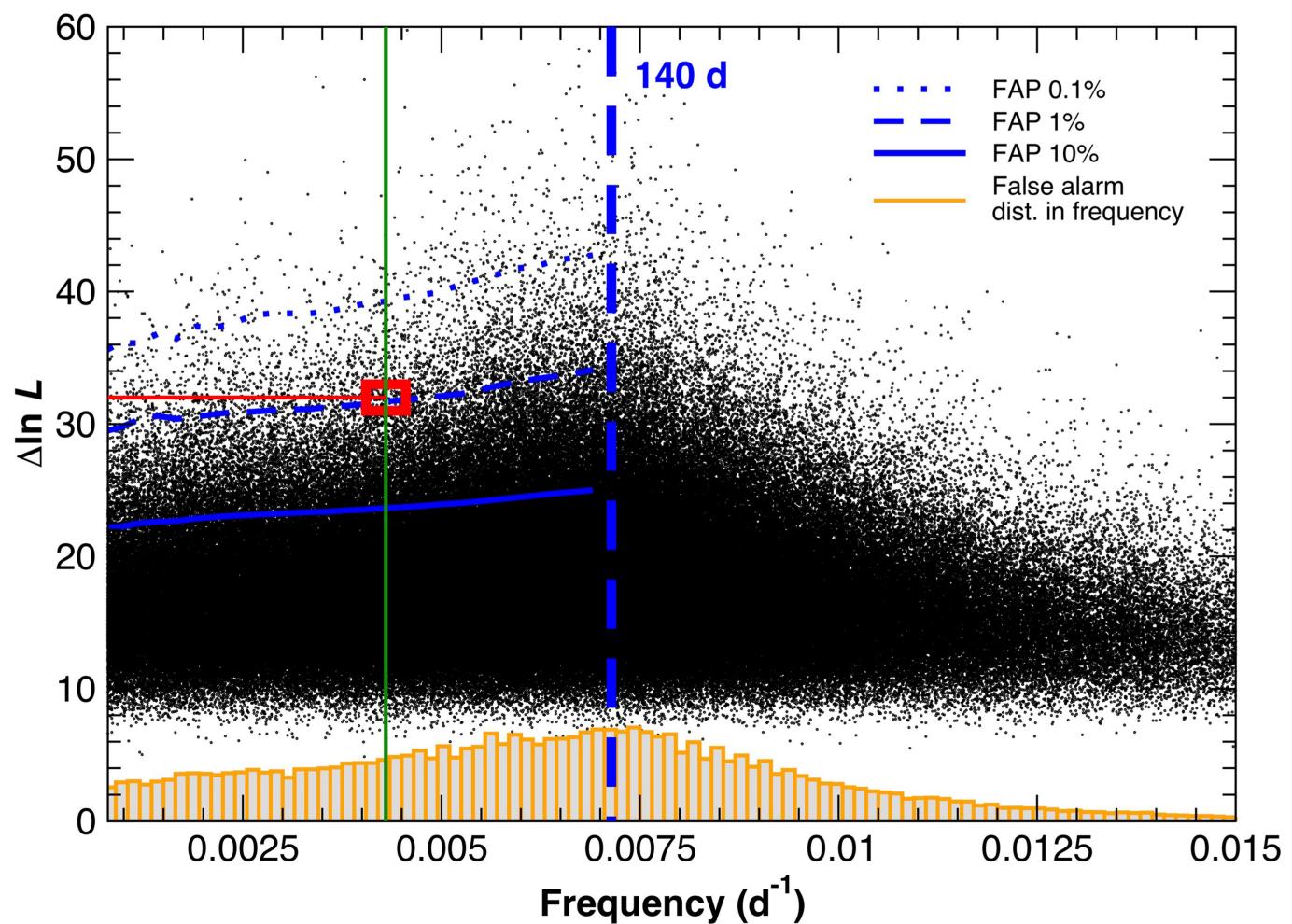
Extended Data Fig. 3 | Propagation of astrometric errors to radial velocity systematics. **a**, Spurious radial-velocity effect ΔRV that would be caused by offsets with respect to the catalogue coordinates (black and red) and proper motions (green and blue). **b**, Illustration of the radial-velocity effect caused by an offset in the parallax with respect to the catalogue

value. The uncertainties of the astrometric parameters for Barnard's star from the Hipparcos catalogue were used in the barycentric corrections, and are approximately 10 times smaller than the values used in this plot (15 mas in position, 1.5 mas yr⁻¹ in proper motion and 1.5 mas in parallax), implying that catalogue errors introduce undetectable signals.



Extended Data Fig. 4 | Effect of Gaussian-process modelling applied to synthetic and real data. Blue squares represent the improvement in the log-likelihood using a Gaussian process to model the correlated noise when trying to detect a first signal. The same procedure is applied to simulated observations generated with white noise and a sinusoidal signal consistent with the parameters of the candidate planet (red circles). Even in absence of true correlated noise, the Gaussian process absorbs a substantial amount of significance ($\Delta \ln L \approx 30$ for this selection of kernel parameters). The adopted kernel is a damped SHO, with a

damping timescale of $\tau = P_{\text{life}} = 100$ days, and each point corresponds to different values for the oscillator frequency ν (x axis). The PSD of an SHO kernel with $\nu = 140^{-1} \text{ d}^{-1}$ and $\tau = 100$ days is depicted as a black line. The greatest reduction in significance occurs when the trial frequency approaches that of the oscillator, but this reduction in significance extends out to a broad range of frequencies, therefore acting as a filter. The period of the candidate planet is marked with a vertical dashed line, and the likely rotation period derived from stellar activity is marked with a vertical dotted line.



Extended Data Fig. 5 | Distribution of empirical false alarms from synthetic observations with correlated noise. These simulations were obtained by generating synthetic observations following kernels derived from the observations, and then fitted to moving-average models. The resulting distribution of false alarms shows a clear excess around the measured rotation period of the star (vertical dashed blue line) and at low frequencies (long periods), owing to the use of the free offsets in the model (left of the rotation period). The empirical FAP was computed by

counting the number of false alarms in the interval $\Delta \ln L \in [32, \infty)$ and frequency $\in [0, 1/230]$ (left of the green line and above the red line) and dividing by total number of false alarms in the same frequency interval (left of the green line). Empirical FAP thresholds of 10%, 1% and 0.1% are shown for reference. The candidate signal under discussion is shown as a red square and has an empirical FAP of about 0.8%. The orange histogram at the bottom shows the distribution of false alarms in frequency (arbitrary normalization).

Extended Data Table 1 | Log of observations of Barnard's star

Instrument	Calib. method	Time	Epochs	Program ID	PI/Group
Keck/HIRES	Iodine	06/1997–08/2013	186	†	Vogt, Butler, Marcy, Fischer, Borucki, Lissauer, Johnson (and several more with <10 obs)
VLT/UVES	Iodine	04/2000–10/2006	75	65.L-0428 66.C-0446 267.C-5700 68.C-0415 69.C-0722 70.C-0044 71.C-0498 072.C0495 173.C-0606 078.C-0829	UVES survey; Kürster
ESO/HARPSpre	Hollow-cathode lamp	04/2007–05/2013	118	072.C-0488 183.C-0437 191.C-0505	Mayor, Bonfils, Anglada-Escudé
Magellan/PFS	Iodine	08/2011–08/2016	39	Carnegie-California survey	Crane, Butler, Shectman, Thompson
APF	Iodine	07/2013–07/2016	43	LCES/APF planet survey	Vogt, Butler (and several programmes)
HARPS-N	Hollow-cathode lamp	07/2014–10/2017	40	CAT14A_43 A27CAT_83 CAT13B_136 CAT16A_99 CAT16A_109 CAT17A_38 CAT17A_58 CAT17B_140	Amado, Rebolo, González Hernández, Berdiñas
CARMENES	Hollow-cathode lamp	02/2016–11/2017	201	CARMENES GTO survey	CARMENES consortium
ESO/HARPSpost	Hollow-cathode lamp	07/2017–09/2017	69	099.C-0880	Anglada-Escudé/RedDots

In the case of ESO/HARPS, the 'pre' and 'post' tags indicate data obtained before and after a hardware upgrade in June 2015. A secular acceleration term of $4.497 \text{ m s}^{-1} \text{ yr}^{-1}$ due to change in perspective over time⁴ was removed from all datasets when applying the barycentric correction to the raw Doppler measurements. The final column lists the Principal Investigators of the proposals that obtained the relevant measurements.

†H7aH, K01H, N02H, N03H, N05H, N06H, N10H, N12H, N14H, N15H, N19H, N20H, N22H, N24H, N28H, N31H, N50H, N59H, U01H, U05H, U07H, U08H, U10H, U11H, U12H, U66H, H38bH, A264Hr, A285Hr, A288Hr, C110Hr, C168Hr, C169Hr, C199Hr, C202Hr, C205Hr, C232Hr, C240Hr, C275Hr, C332Hr, H174Hr, H218Hr, H238Hr, H244Hr, H257Hr, H305Hr, N007Hr, N014Hr, N023Hr, N024Hr, N054Hr, N085Hr, N086Hr, N095Hr, N108Hr, N118Hr, N125Hr, N129Hr, N131Hr, N134Hr, N136Hr, N141Hr, N145Hr, N148Hr, N157Hr, N168Hr, U009Hr, U014Hr, U023Hr, U026Hr, U027Hr, U030Hr, U052Hr, U058Hr, U064Hr, U077Hr, U078Hr, U082Hr, U084Hr, U115Hr, U131Hr, U142Hr, Y013Hr, Y065Hr, Y292Hr.

Extended Data Table 2 | Additional fit parameters and fit results

Dataset	Jitter (m s^{-1})	γ (m s^{-1})
Keck/HIRES	$2.28^{+0.19}_{-0.18}$	$1.26^{+0.38}_{-0.32}$
VLT/UVES	$2.42^{+0.25}_{-0.22}$	$3.83^{+0.58}_{-0.57}$
ESO/HARPSpre	0.92 ± 0.14	$0.97^{+0.36}_{-0.27}$
Magellan/PFS	$0.96^{+0.37}_{-0.41}$	$1.76^{+0.41}_{-0.37}$
APF	$2.78^{+0.51}_{-0.44}$	$2.16^{+0.65}_{-0.63}$
HARPS-N	$1.45^{+0.27}_{-0.23}$	$1.37^{+0.65}_{-0.63}$
CARMENES	$1.76^{+0.15}_{-0.14}$	1.55 ± 0.65
ESO/HARPSpost	$1.16^{+0.19}_{-0.18}$	1.46 ± 0.69

The individual zero points γ and jitter terms are optimized for each dataset by maximizing the likelihood function. The model also included a parameter representing a global linear radial velocity trend over time, for which the optimization process yielded a best-fitting value of $+0.33 \pm 0.07 \text{ m s}^{-1} \text{ yr}^{-1}$. The original individual datasets were previously shifted to have null relative offsets in the overlapping regions (see Extended Data Table 3) and referred to the zero-point level of the Keck/HIRES dataset. This implies that the optimized γ parameters in the table are not totally arbitrary but expected to be relatively similar. The parameters and their uncertainties are determined by calculating the median values and 68% credibility intervals of the distribution that results from the MCMC run.

Extended Data Table 3 | Zero-point offsets between overlapping radial-velocity datasets from different instruments

Datasets	Window	Measur.	Diff. (manual-optimized)
	size (\pm days)	used	(m s^{-1})
UVES-HIRES	10	28	2.53 ± 0.65
HARPSpre-HIRES	10	291	-0.29 ± 0.31
PFS-HIRES	10	130	0.49 ± 0.49
APF-PFS	10	17	0.37 ± 0.85
HARPSpost-CARMENES	2	161	-0.09 ± 0.33
HARPS-N-CARMENES	2	75	-0.18 ± 0.39
Set1-Set2	8	14	-0.24 ± 0.52

Manual offsets are calculated from common regions of pairs of datasets for window sizes selected to ensure sufficient statistics and consistency in the case of three-way overlap. The last column lists the difference between the zero points calculated manually and those resulting from the global optimization, demonstrating general good agreement (values compatible with zero), except for the UVES dataset. Also, two distinct time regions are identified in the data and can be compared. Set 1 includes data from HIRES, UVES, HARPSpre, APF and PFS. Set 2 contains data from CARMENES, HARPS-N and HARPSpost. The relative zero point between these two sets is poorly defined because of very limited overlap, but the consistency between the manual and optimized values is very good. All errors correspond to 1σ .

Emergence of multi-body interactions in a fermionic lattice clock

A. Goban^{1,2,7*}, R. B. Hutson^{1,2,7}, G. E. Marti^{1,2,6}, S. L. Campbell^{1,2,4,5}, M. A. Perlin^{1,2}, P. S. Julienne³, J. P. D’Incao^{1,2}, A. M. Rey^{1,2} & J. Ye^{1,2*}

Alkaline-earth atoms have metastable ‘clock’ states with minute-long optical lifetimes, high-spin nuclei and $SU(N)$ -symmetric interactions, making them powerful platforms for atomic clocks¹, quantum information processing² and quantum simulation³. Few-particle systems of such atoms provide opportunities to observe the emergence of complex many-body phenomena with increasing system size⁴. Multi-body interactions among particles are emergent phenomena, which cannot be broken down into sums over underlying pairwise interactions. They could potentially be used to create exotic states of quantum matter^{5,6}, but have yet to be explored in ultracold fermions. Here we create arrays of isolated few-body systems in an optical clock based on a three-dimensional lattice of fermionic ^{87}Sr atoms. We use high-resolution clock spectroscopy to directly observe the onset of elastic and inelastic multi-body interactions among atoms. We measure the frequency shifts of the clock transition for varying numbers of atoms per lattice site, from $n = 1$ to $n = 5$, and observe nonlinear interaction shifts characteristic of elastic multi-body effects. These measurements, combined with theory, elucidate an emergence of $SU(N)$ -symmetric multi-body interactions, which are unique to fermionic alkaline-earth atoms. To study inelastic multi-body effects, we use these frequency shifts to isolate n -occupied sites in the lattice and measure the corresponding lifetimes of the clock states. This allows us to access the short-range few-body physics without experiencing the systematic effects that are encountered in a bulk gas. The lifetimes that we measure in the isolated few-body systems agree very well with numerical predictions based on a simple model for the interatomic potential, suggesting a universality in ultracold collisions. By connecting these few-body systems through tunnelling, the favourable energy and timescales of the interactions will allow our system to be used for studies of high-spin quantum magnetism^{7,8} and the Kondo effect^{3,9}.

Fermionic alkaline-earth and alkaline-earth-like atoms have $^1\text{S}_0$ ground clock states and metastable $^3\text{P}_0$ excited clock states ($^3\text{P}_0$ has a lifetime of roughly 160 s in ^{87}Sr), which provide two (electronic) orbital degrees of freedom that are largely decoupled from the nuclear spin I . This decoupling gives rise to orbital, $SU(N=2I+1)$ -symmetric, two-body interactions in which the s -wave and p -wave scattering parameters are independent of the nuclear spin state^{3,8}. This degeneracy can be quite large ($I=9/2$ for ^{87}Sr), enabling studies of quantum states of matter with no direct analogues in nature, such as the $SU(N)$ Mott insulator^{3,10,11}. Two-orbital, $SU(N)$ -symmetric interactions were first observed directly using clock spectroscopy^{7,12,13} and have since provided new opportunities for studying strongly interacting Fermi gases^{14,15} and the Kondo lattice model⁹.

Whereas particles microscopically interact in a pairwise manner, multi-body interactions can emerge in a low-energy effective field theory in which fluctuations beyond some length or momentum scale are integrated out. Examples of such multi-body interactions include three-nucleon forces¹⁶ and some fractional quantum Hall states⁵. Multi-body

interactions have been predicted to arise in various optical lattice experiments^{6,17} and have been observed in bosonic systems^{18,19}. Although a single impurity interacting with a few identical fermions has been studied⁴, multi-body interactions in high-spin fermions have yet to be explored.

In ultracold gases, the effects of multi-body interactions have been explored extensively in the context of three-body recombination processes²⁰, including in studies of exotic Efimov states and of other forms of universality associated with long-range interactions^{21,22}. However, comparison to theory is often difficult owing to the bulk-gas nature of these experiments. Improved control and understanding of the external degrees of freedom of atoms is crucial for testing theoretical models of ultracold collisions^{23,24}.

Here we study the emergence of multi-body interactions by combining isolated few-body systems in an optical lattice with high-resolution clock spectroscopy. In our experiment, the ultracold gas is prepared similarly to previous work^{25,26}. In summary, we prepare a ten-spin-component Fermi degenerate gas, with atoms distributed equally among all nuclear spin states. We typically produce 10^3 – 10^4 atoms per nuclear spin state at a temperature $T = 10$ – 20 nK = $0.1T_{\text{F}}$, where T_{F} is the Fermi temperature. The gas is loaded into a nearly isotropic, three-dimensional optical lattice, in which the geometric mean of the trap depths for the three lattice beams \mathcal{U} varies from $30E_{\text{rec}}$ to $80E_{\text{rec}}$, where $E_{\text{rec}} = h \times 3.5$ kHz is the recoil energy of a lattice photon and h is the Planck constant. At these trap depths, there is negligible tunnelling between neighbouring sites over the timescale of the experiment.

As depicted in Fig. 1a, for atoms in doubly occupied sites, a π -polarized clock photon resonantly couples the ground state $|\text{gg}\rangle$ to the orbitally symmetric (or orbitally antisymmetric) excited state $|\text{eg}^+\rangle$ (or $|\text{eg}^-\rangle$) upon matching the detuning $(E_{\text{eg}^+} - E_{\text{gg}})/h$ (or $(E_{\text{eg}^-} - E_{\text{gg}})/h$) at zero magnetic field. Here, ‘g’ and ‘e’ represent the ground clock state $^1\text{S}_0$ and the excited clock state $^3\text{P}_0$, respectively, and E_X is the on-site interaction energy for $X \in \{\text{gg}, \text{eg}^\pm\}$. Similarly, for sites with $n \geq 3$, the ground state $|\text{g}\dots\rangle$ can be driven to the orbitally symmetric state $|\text{eg}\dots^+\rangle$ or to the state $|\text{eg}\dots^-\rangle$, for which the orbital and nuclear spin degrees of freedom are not separable. The π -polarized clock light preserves the initial distribution of the nuclear spin states.

We spatially resolve the spectroscopic signal using absorption imaging²⁶ and the readout scheme presented in Fig. 1b. We measure the differential interaction energies and the spatial distributions of each occupation number^{27,28}. In Fig. 2a we show sample spectra of a ten-spin-component Fermi gas using 20-ms clock pulses from a 26-mHz-line width ultrastable laser. For each occupation number n , there is a pair of single-excitation resonances, labelled n^\pm , which correspond to the two sets of final states $|\text{eg}\dots^\pm\rangle$. $SU(N)$ symmetry and fermionic antisymmetrization dictate that only two eigenenergies appear for each n -atom sample (see Supplementary Information).

In Fig. 2b we show the column density of different occupation numbers for a sample of 2×10^5 atoms. The shells of decreasing size with

¹JILA, National Institute of Standards and Technology, University of Colorado, Boulder, CO, USA. ²Department of Physics, University of Colorado, Boulder, CO, USA. ³Joint Quantum Institute, NIST, University of Maryland, Gaithersburg, MD, USA. ⁴Present address: Molecular Biophysics and Integrated Bioimaging, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ⁵Present address: Department of Physics, University of California, Berkeley, CA, USA. ⁶Present address: Department of Molecular and Cellular Physiology, Stanford University, Stanford, CA, USA. ⁷These authors contributed equally: A. Goban, R. B. Hutson. *e-mail: Akihisa.Goban@jila.colorado.edu; Ye@jila.colorado.edu

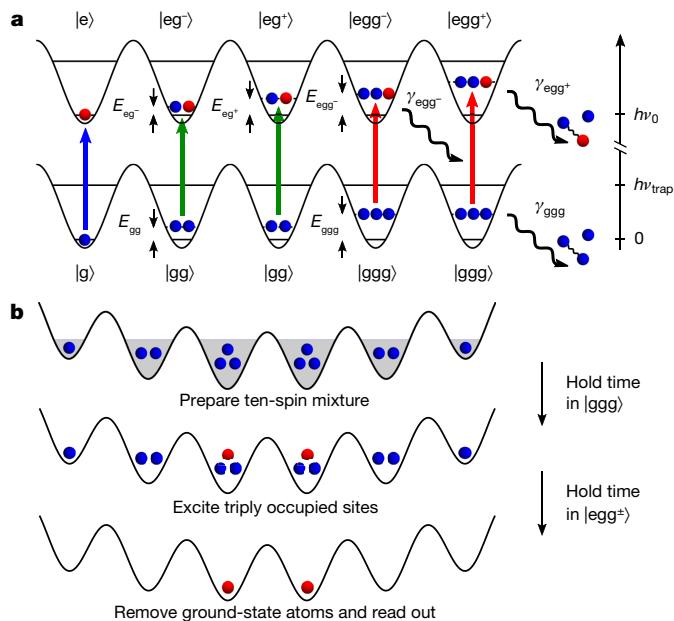


Fig. 1 | Two-orbital interactions in a three-dimensional lattice and experimental sequence. **a**, One to three ^{87}Sr atoms (blue and red circles for electronic ground- and excited-state atoms, respectively) occupy the lowest motional state of a lattice site, with corresponding on-site energies E_X . We use a state-independent lattice that operates at the ‘magic’ wavelength, at which the polarizabilities of the ground and excited states are identical. In a deep three-dimensional lattice, each site can be regarded as an isolated few-body system with a trap frequency ν_{trap} . A clock photon resonantly couples the ground state $|\text{g}\cdots\rangle$ to the single-excitation manifold $|\text{eg}\cdots^{\pm}\rangle$, leading to a spectroscopic shift from the bare resonance frequency $\nu_0 \approx 429$ THz. Multi-body interactions manifest in sites with three or more atoms, both in the observed clock shifts and in their decay (black squiggly arrows) into a diatomic molecule plus a free atom at a rate γ_X . **b**, Experimental sequence for imaging triply occupied sites. A ten-nuclear-spin mixture is loaded into a three-dimensional optical lattice. A clock pulse resonantly drives triply occupied sites $|\text{ggg}\rangle$ to an excited state $|\text{egg}^{\pm}\rangle$. After all atoms in the ground state are removed, the remaining atoms, in the excited state, are read out using absorption imaging. Three-body decay rates are measured by adding a hold time before (for γ_{ggg}) or after (for $\gamma_{\text{egg}^{\pm}}$) applying the clock pulse.

increasing occupation number are a result of balancing the external confinement generated by Gaussian lattice beams with the on-site interaction energies. As observed for small n , larger clouds of atoms extend over areas where the trapping frequencies are relatively lower, resulting in smaller on-site interaction energies. To eliminate a possible systematic shift from the changing cloud size, we adjust the final evaporation point to maximize the central density of the desired occupation number and measure the spectroscopic response in only the central $4 \mu\text{m} \times 4 \mu\text{m} \times 2 \mu\text{m}$ region of the trap. The vertical plane is selected by loading the lattice from a trap that is tightly confining against gravity, loading only a $2\text{-}\mu\text{m}$ -thick vertical region. Spatial selection in the horizontal plane is performed by spatially filtering the images, measuring the response from only the central region of the lattice. The trap depth in the central region of the lattice is calibrated via motional sideband spectroscopy of an $n=1$ sample with the same spatial selection.

To investigate multi-body interactions in multiply occupied sites, we consider the case of two interacting fermionic atoms, each with two internal degrees of freedom: an electronic orbital, $x \in \{\text{g}, \text{e}\}$, and a nuclear spin sublevel, $m \in \{-I, -I+1, \dots, I\}$. The interactions depend on only the electronic degree of freedom, so all s -wave scattering processes are parameterized by four scattering lengths a_X , with $X \in \{\text{gg}, \text{eg}^+, \text{eg}^-, \text{ee}\}$, resulting in $\text{SU}(N)$ -symmetric properties of the system. Here, ‘+’ (‘-’) denotes a symmetric (antisymmetric) superposition of the electronic orbitals. In our experiments, atoms are trapped in the motional ground states of deep lattice sites with a

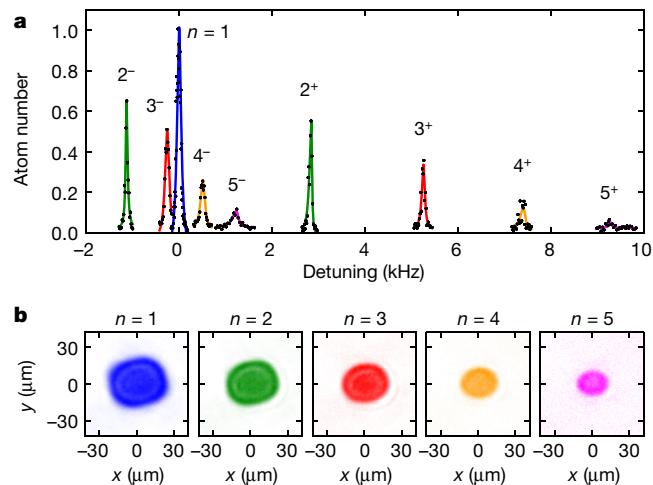


Fig. 2 | Clock spectroscopy of a ten-component Fermi gas in a three-dimensional lattice. **a**, Overlaid clock spectra for occupation numbers $n=1, \dots, 5$ at a mean trap depth of $U=54E_{\text{rec}}$ ($\nu_{\text{trap}}=51$ kHz). The labels n^{\pm} denote the excitation $|\text{eg}\cdots^{\pm}\rangle$ for n -occupied sites. For large occupations, the line shapes become asymmetric owing to the inhomogeneity of the trap depth. The solid lines are fits used to determine the resonance frequencies (see Methods). The detunings are given relative to the resonance of the clock transition for singly occupied sites (blue) and the atom number is normalized such that the peak height for singly occupied sites is unity. Each data point is the result of a single experimental cycle. **b**, Column densities of different occupation numbers for a sample of 2×10^5 atoms. The absorption images for different occupation numbers were obtained according to the procedure in Fig. 1b, by first exciting the symmetric resonances. Each image is averaged over 20 experimental cycles.

single-particle Wannier function $\phi_0(\mathbf{r})$, localized to a characteristic length scale l_0 . Because all atoms are in the motional ground states, the Pauli exclusion principle requires that atoms with the same orbital state x have different nuclear spins m . Here, we consider the case in which each atom is in a different spin state.

In the limit of weak interactions ($l_0 \gg |a_X|$), the pairwise interaction energy can be expressed as

$$U_X^{(2)} = \frac{4\pi\hbar^2}{m_a} a_X \int |\phi_0(\mathbf{r})|^4 d^3\mathbf{r}$$

where m_a is the atomic mass and $\hbar = h/(2\pi)$. In this regime, the on-site many-body Hamiltonian is

$$H = \sum_{m \neq m'} \left[\frac{U_{\text{gg}}^{(2)}}{2} n_{\text{g},m} n_{\text{g},m'} + \frac{U_{\text{ee}}^{(2)}}{2} n_{\text{e},m} n_{\text{e},m'} + V_{\text{ex}}^{(2)} c_{\text{e},m}^{\dagger} c_{\text{g},m}^{\dagger} c_{\text{e},m'} c_{\text{g},m'} \right] \quad (1)$$

where $c_{x,m}^{\dagger}$ ($c_{x,m}$) creates (destroys) an atom in orbital $x \in \{\text{e}, \text{g}\}$ with spin m , and $n_{x,m} = c_{x,m}^{\dagger} c_{x,m}$. The direct and exchange interaction energies are $V_{\text{eg}}^{(2)} = (U_{\text{gg}}^{(2)} + U_{\text{ee}}^{(2)})/2$ and $V_{\text{ex}}^{(2)} = (U_{\text{gg}}^{(2)} - U_{\text{ee}}^{(2)})/2$, respectively. For a lattice site occupied by n atoms, the Hamiltonian in equation (1) has a ground state $|\text{g}\cdots\rangle$ with a corresponding eigenenergy $E_{\text{g}\cdots}$ and a manifold of singly excited states with two distinct eigenenergies $E_{\text{eg}^{\pm}}$, one for the orbitally symmetric states $|\text{eg}\cdots^+\rangle$ and the other for the $(n-1)$ -fold-degenerate states $|\text{eg}\cdots^-\rangle$. The states $|\text{g}\cdots\rangle$ and $|\text{eg}\cdots^+\rangle$ are each symmetric in their orbital degree of freedom and, as such, fermionic statistics requires their nuclear spin degree of freedom to form an $\text{SU}(N)$ singlet. The orbital degree of freedom of $|\text{eg}\cdots^+\rangle$ is an n -body W state, which constitutes an important resource for quantum information processing and quantum communications protocols²⁹. For $n \geq 3$, the $|\text{eg}\cdots^-\rangle$ states are highly entangled between their orbital and nuclear spin degrees of freedom such that each degree of freedom is of mixed symmetry (see Supplementary Information).

Table 1 | s-Wave scattering lengths and three-body loss coefficients

Channel, X	s-Wave scattering lengths, $a_X (a_0)$	Two-body loss coefficients, $\beta_X (10^{-16} \text{ cm}^3 \text{ s}^{-1})$
	Measured	Calculated
Three-body loss coefficients, $\beta_X (10^{-30} \text{ cm}^3 \text{ s}^{-1})$		
gg	2.0(0.2)	1.7
eg [−]	25(1)	26
eg ⁺	15(1)	8.0

The scattering length of ground states $a_{gg} = 96.2(0.1)a_0$ is determined from photoassociation spectroscopy³⁰; all the other values are from this work. The measured elastic s-wave scattering lengths a_X are consistent with previously reported values⁷, with an improvement by a factor of ten in the uncertainty for $X = \text{eg}^+$. The subscripts 'stat' and 'sys' denote the statistical and systematic uncertainty, respectively (see Supplementary Information). The two-body loss coefficients β_X are upper bounds, limited by the excited-state lifetime. The measured three-body loss coefficients are in good agreement with the ones calculated using a universal van der Waals model.

For tighter confinement and stronger on-site interactions—that is, if a_X/l_0 is not negligible—corrections to equation (1) become increasingly important. The increased interaction energy facilitates off-resonant transitions to higher motional states. Equivalently, the spatial wavefunction $\phi_0(\mathbf{r})$ becomes dependent on the number of atoms per site and their configuration. This effect can be captured by a lowest-band effective Hamiltonian where the higher motional states are integrated out, with two consequences: (i) the two-body interaction energies are characterized by an in-trap scattering length, rescaled from the free-space one; and (ii) the total interaction energy for $n \geq 3$ atoms cannot be broken down into a sum over pairs of atoms¹⁷, leading to effective multi-body interactions. Considering at most one atom in the excited state, equation (1) must be modified to include multi-body corrections:

$$H' = \sum_{m=m' \neq m''} \left[\frac{U_{ggg}^{(3)}}{6} n_{g,m} n_{g,m'} n_{g,m''} + \frac{V^{(3)}}{2} n_{e,m} n_{g,m'} n_{g,m''} \right. \\ \left. + \frac{V_{\text{ex}}^{(3)}}{2} c_{e,m}^\dagger c_{g,m}^\dagger c_{e,m} c_{g,m'} n_{g,m''} \right] + \mathcal{O}(n^4) \quad (2)$$

where $U_{ggg}^{(3)}$, $V^{(3)}$ and $V_{\text{ex}}^{(3)}$ are the effective three-body ground-state, direct and exchange interaction energies. Owing to the SU(N) symmetry, H' has the same eigenstates as H , but with modified n -body eigenenergies (see Supplementary Information). These multi-body interactions can be probed by spectroscopically addressing lattice sites with different occupation numbers.

To extract the multi-body effects from equation (2), we measure the frequency shifts $(E_{\text{eg}^+} - E_{\text{gg}})/h$ for various mean trap depths. By incorporating the corrections due to lattice confinement with a previous measurement of the ground-state scattering length $a_{gg} = 96.2(0.1)a_0$, where a_0 is the Bohr radius³⁰, we extract the free-space scattering lengths a_{eg^\pm} , shown in Table 1 (see Supplementary Information).

Multi-body interactions occur only at sites with three or more atoms and cause frequency shifts that are nonlinear in the occupation number n . The measured clock shifts of the $|\text{eg}^{\pm}\rangle$ branches are shown as red points in Fig. 3a, b. These clock shifts deviate from the values expected from equation (1) (blue triangles), which are proportional to the occupation number n , and are consistent with the three-body corrections in equation (2) (black squares). These higher-order contributions (Fig. 3c, d) can be intuitively interpreted as broadening of the wavefunction, lowering the magnitude of the overall interaction energy. From a variational calculation, we find that the wavefunction for $n = 5$ atoms is broadened by about 8% relative to a non-interacting wavefunction.

Multi-body effects also appear in these few-body systems as three-body recombination loss. These losses occur when three atoms recombine to form a deeply bound diatomic molecule and a free atom, each

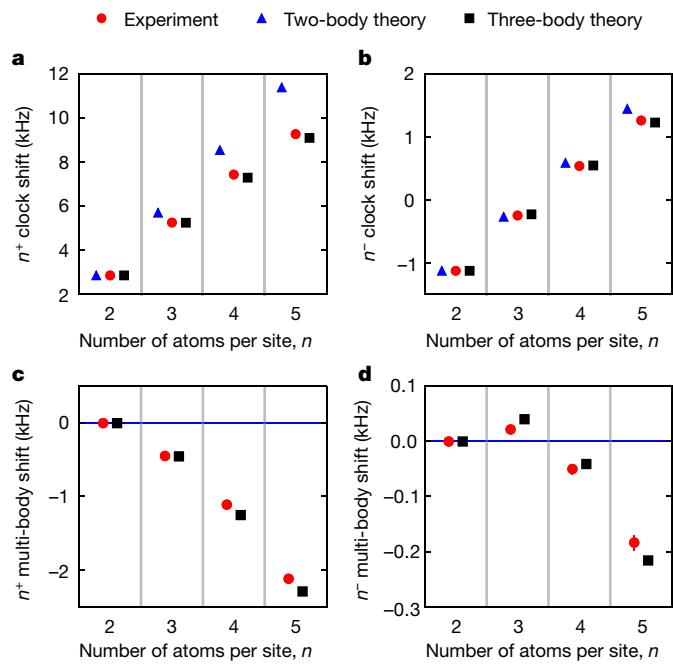


Fig. 3 | Effective multi-body clock shifts. **a**, Clock shifts of $|\text{eg}^{\pm}\rangle$ from $n = 2, \dots, 5$ at a mean trap depth of $\mathcal{U} = 54E_{\text{rec}}$ ($\nu_{\text{trap}} = 51 \text{ kHz}$). Effective multi-body interactions are observed in the experimental data (red circles) as a deviation from the two-body prediction (blue triangles). The calculated shifts from an effective Hamiltonian including three-body interactions (see text and Supplementary Information) are shown as black squares. The points at each occupation number n are offset horizontally for clarity. The uncertainties of the experimental data are smaller than the size of the symbols. **b**, Clock shifts of $|\text{eg}^{\pm}\rangle$ under the same conditions as in **a**. The two-body theory shows smaller deviations from the measured shifts at $n = 3$ and $n = 4$, owing to a near cancellation of the three-body shifts between $|\text{g}\cdots\rangle$ and $|\text{eg}^{\pm}\rangle$. **c, d**, Multi-body interaction shifts, which correspond to the data from **a** and **b** with two-body contributions subtracted. All error bars are 1 s.e., determined from fits of the resonance position.

carrying enough energy to eject it from the trap²⁴. We selectively determine the lifetime of a given n -atom $|\text{g}\cdots\rangle$ state by holding atoms in a deep lattice for a variable time, then resonantly driving the transition $|\text{g}\cdots\rangle \rightarrow |\text{eg}^{\pm}\rangle$ to spectroscopically address only the n -atom sites, and finally measuring the excited-state atom population after removing the ground-state atoms, as illustrated in Fig. 1. Similarly, the loss rate of the $|\text{eg}^{\pm}\rangle$ state is determined by first driving the transition $|\text{g}\cdots\rangle \rightarrow |\text{eg}^{\pm}\rangle$, and then holding for a variable time before removing the ground-state atoms. Because this procedure measures the probability for the n -atom system to remain in its initial state, the spectroscopic signal decays in a simple and robust form as a single exponential, which we fit to extract a 1/e lifetime. This analysis is much simpler than that for bulk-gas experiments, for which decay curves must be fitted with multiple rate constants corresponding to one-, two- and three-body losses^{31,32}.

To disentangle these multi-body effects from inelastic two-body collisions, we measure the ground- and excited-state lifetimes of the one- and two-atom sites. Whereas we observe a vacuum-limited 1/e lifetime of around 100 s for the $|\text{g}\rangle$ and $|\text{gg}\rangle$ states, off-resonant Raman scattering from the optical lattice light causes a decay of the single-atom excited state, $|\text{e}\rangle \rightarrow |\text{g}\rangle$, with a time constant of 9.6(0.4) s at a mean trap depth of $\mathcal{U} = 73E_{\text{rec}}$ (R.B.H. et al., manuscript in preparation). Note that all errors for the lifetimes are 1 s.e. from exponential fits. At this same trap depth, we find the lifetimes τ_{eg^\pm} of the $|\text{eg}^+\rangle$ and $|\text{eg}^-\rangle$ states to be 5.1(0.7) s and 6.1(0.7) s, respectively. Such two-body lifetimes can be related to a two-body loss coefficient β_{eg^\pm} via the expression

$$\tau_{\text{eg}^\pm}^{-1} = \beta_{\text{eg}^\pm} \int |\phi_0(\mathbf{r})|^4 d^3\mathbf{r}$$

However, because the two-body lifetimes are only slightly shorter than that of a single excited atom, we can determine only upper limits: $\beta_{\text{eg}^+} \leq 2.5(0.3) \times 10^{-16} \text{ cm}^3 \text{ s}^{-1}$ and $\beta_{\text{eg}^-} \leq 2.1(0.2) \times 10^{-16} \text{ cm}^3 \text{ s}^{-1}$. The measured lifetimes of the three-atom states, τ_X for $X \in \{\text{ggg}, \text{egg}^+, \text{egg}^-\}$, at various mean trap depths are shown in Fig. 4a. These multi-body decays all occur on timescales much shorter than those of one- and two-body losses. Furthermore, the excited states are observed to decay faster than the ground state by approximately an order of magnitude. We attribute this to the increased number of molecular decay channels after replacing a ground-state atom with a distinguishable excited-state atom. Table 1 shows the density-independent three-body loss coefficient β_X extracted from these measurements via the expression

$$\tau_X^{-1} = \beta_X \int |\phi_0(\mathbf{r})|^6 d^3r$$

Next, we compare the measured three-body lifetimes to a model in which atoms interact by pairwise, additive, long-range van der Waals potentials joined at shorter range to a pseudopotential that is adjusted to yield, in each case, the measured two-body scattering lengths given in Table 1^{7,24,33}. Numerically solving the three-body Schrödinger equation yields the frequency shifts and decay lifetimes for three atoms confined in a harmonic trap (see Supplementary Information). We increase the number of bound states in each pairwise potential until all results converge to less than 10%. As shown in Fig. 4a, the calculated lifetimes (open circles) are remarkably close to the measured lifetimes (filled circles) given the simplicity of our universal van der Waals model, which has no fit parameters. Whereas our results for $|\text{ggg}\rangle$ and $|\text{egg}^-\rangle$ agree with the observed lifetimes to within 15% or less, the results for $|\text{egg}^+\rangle$ overestimate the lifetimes by about 50%, most probably owing to the fact that for this state our model does not allow for decay into all possible diatomic molecular states (see Supplementary Information). As a sanity check, the frequency shifts produced by this model agree with the measurements shown in Fig. 3a, b to within 10%, despite assuming a harmonic trap potential.

We extract three-body loss coefficients β_X from the calculated lifetimes using the same procedure as for the experimental results shown in Table 1. The good agreement between the universal van der Waals model and our ^{87}Sr lattice experiment is in sharp contrast to the disagreement, by a factor of 2–4, for bulk-gas ^{87}Rb experiments. The universal van der Waals model of ^{87}Rb gives a three-body loss rate coefficient of²⁴ $1.0 \times 10^{-29} \text{ cm}^6 \text{ s}^{-1}$, in contrast to the measured value of³¹ $4.3(1.8) \times 10^{-29} \text{ cm}^6 \text{ s}^{-1}$. This scenario suggests that a lattice experiment with ^{87}Rb could greatly decrease the uncertainty in the ^{87}Rb three-body recombination loss coefficient and provide a better test of the theory for that system.

Finally, we study the dependence of the lifetimes on the occupation number. In Fig. 4b we show a

$$\tau_{g\dots}^{-1} = \tau_{\text{ggg}}^{-1} \binom{n}{3}$$

scaling of the lifetime of the n -body ground state for $n \geq 3$, which suggests that three-body loss remains the dominant mechanism. The lifetimes of the n -atom excited states, along with their expected scalings from counting the number of three-body loss channels, are shown in Fig. 4c (see Supplementary Information). These relatively long lifetimes are promising for future experiments involving coupled wells with large occupation numbers.

In conclusion, we have demonstrated two manifestations of multi-body interactions arising from pairwise interactions in few-body systems of fermions. Our spectroscopic technique, along with spatially resolved readout, enables efficient isolation of few-body systems, which prove to be ideal for observing multi-body effects. It also provides a simple way to create the highly entangled and long-lived states $|\text{eg}^{\dots\pm}\rangle$, a useful resource for quantum information processing²⁹. The few-body systems enable precise measurements of the a_{eg^\pm} scattering lengths and the three-body loss rates which agree with the universal van der Waals

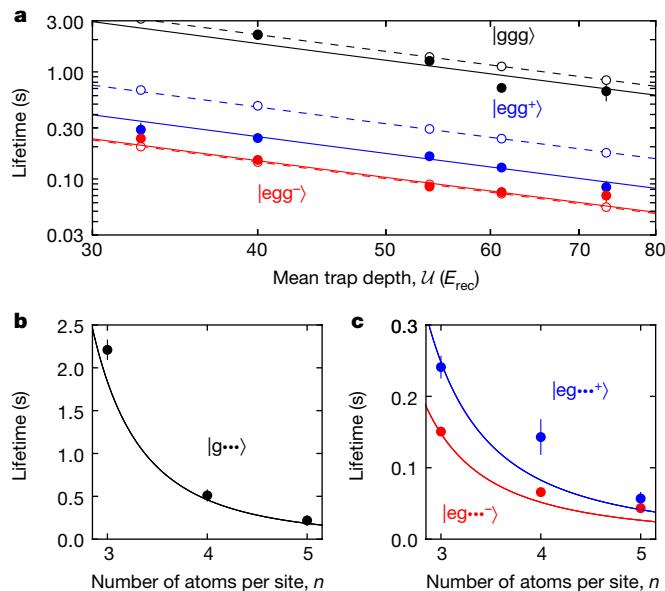


Fig. 4 | Three-body loss rate and occupation-number-dependent lifetime. **a**, Dependence of the $n = 3$ lifetimes on the mean trap depth for the $|\text{ggg}\rangle$ (black), $|\text{egg}^+\rangle$ (blue) and $|\text{egg}^-\rangle$ (red) states. The measured lifetimes (filled circles) are close to the ones calculated from a universal van der Waals model (open circles; see text and Supplementary Information). From the fits shown as solid (dashed) lines, we extract the three-body loss coefficients β_X for the measured (calculated) lifetimes, which are summarized in Table 1. The lifetime of $|\text{ggg}\rangle$ is ten times that of $|\text{egg}^\pm\rangle$, because the number of molecular states increases owing to one distinguishable particle in the excited state. **b, c**, Dependence of the lifetimes of the $|\text{g}\dots\rangle$ and $|\text{eg}^{\dots\pm}\rangle$ states on the occupation number n at $U = 40E_{\text{rec}}$. The solid lines are the lifetimes calculated assuming pure three-body losses, with the measured β_{ggg} and β_{egg^\pm} values as input parameters (see text and Supplementary Information). All error bars are 1 s.e., determined from exponential fits.

model. The collisional parameters, in the case of ^{87}Sr , are found to be particularly suitable for studies of two-orbital $\text{SU}(N)$ magnetism, which should arise in the presence of weak tunnelling. These interactions have been predicted to create long-sought states of matter, including valence-bond solids and chiral spin liquids³.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0661-6>.

Received: 15 March 2018; Accepted: 20 August 2018;

Published online 31 October 2018.

1. Ludlow, A. D., Boyd, M. M., Ye, J., Peik, E. & Schmidt, P. O. Optical atomic clocks. *Rev. Mod. Phys.* **87**, 637–701 (2015).
2. Daley, A. J. Quantum computing and quantum simulation with group-II atoms. *Quantum Inform. Process.* **10**, 865–884 (2011).
3. Cazalilla, M. A. & Rey, A. M. Ultracold Fermi gases with emergent $\text{SU}(N)$ symmetry. *Rep. Prog. Phys.* **77**, 124401 (2014).
4. Wenz, A. N. *et al.* From few to many: observing the formation of a Fermi sea one atom at a time. *Science* **342**, 457–460 (2013).
5. Fradkin, E. H., Nayak, C., Tsvelik, A. & Wilczek, F. A Chern-Simons effective field theory for the Pfaffian quantum Hall state. *Nucl. Phys. B* **516**, 704–718 (1998).
6. Büchler, H. P., Micheli, A. & Zoller, P. Three-body interactions with cold polar molecules. *Nat. Phys.* **3**, 726–731 (2007).
7. Zhang, X. *et al.* Spectroscopic observation of $\text{SU}(N)$ -symmetric interactions in Sr orbital magnetism. *Science* **345**, 1467–1473 (2014).
8. Gorshkov, A. V. *et al.* Two-orbital $\text{SU}(N)$ magnetism with ultracold alkaline-earth atoms. *Nat. Phys.* **6**, 289–295 (2010).
9. Rieger, L. *et al.* Localized magnetic moments with tunable spin exchange in a gas of ultracold fermions. *Phys. Rev. Lett.* **120**, 143601 (2018).
10. Taie, S., Yamazaki, R., Sugawa, S. & Takahashi, Y. An $\text{SU}(6)$ Mott insulator of an atomic Fermi gas realized by large-spin Pomeranchuk cooling. *Nat. Phys.* **8**, 825–830 (2012).

11. Hofrichter, C. *et al.* Direct probing of the Mott crossover in the SU(N) Fermi-Hubbard model. *Phys. Rev. X* **6**, 021030 (2016).
12. Scazza, F. *et al.* Observation of two-orbital spin-exchange interactions with ultracold SU(N)-symmetric fermions. *Nat. Phys.* **10**, 779–784 (2014); corrigendum **11**, 514 (2015).
13. Cappellini, G. *et al.* Direct observation of coherent interorbital spin-exchange dynamics. *Phys. Rev. Lett.* **113**, 120402 (2014).
14. Höfer, M. *et al.* Observation of an orbital interaction-induced Feshbach resonance in ^{173}Yb . *Phys. Rev. Lett.* **115**, 265302 (2015).
15. Pagano, G. *et al.* Strongly interacting gas of two-electron fermions at an orbital Feshbach resonance. *Phys. Rev. Lett.* **115**, 265301 (2015).
16. Hammer, H.-W., Nogga, A. & Schwenk, A. Colloquium: Three-body forces: from cold atoms to nuclei. *Rev. Mod. Phys.* **85**, 197–217 (2013).
17. Johnson, P. R., Blume, D., Yin, X. Y., Flynn, W. F. & Tiesinga, E. Effective renormalized multi-body interactions of harmonically confined ultracold neutral bosons. *New J. Phys.* **14**, 053037 (2012); corrigendum **20**, 079501 (2018).
18. Will, S. *et al.* Time-resolved observation of coherent multi-body interactions in quantum phase revivals. *Nature* **465**, 197–201 (2010).
19. Mark, M. J. *et al.* Precision measurements on a tunable Mott insulator of ultracold atoms. *Phys. Rev. Lett.* **107**, 175301 (2011).
20. Ferlaino, F. *et al.* Efimov resonances in ultracold quantum gases. *Few-Body Syst.* **51**, 113–133 (2011).
21. Fletcher, R. J. *et al.* Two- and three-body contacts in the unitary Bose gas. *Science* **355**, 377–380 (2017).
22. Greene, C. H., Giannakeas, P. & Pérez-Ríos, J. Universal few-body physics and cluster formation. *Rev. Mod. Phys.* **89**, 035006 (2017).
23. D’Incao, J. P. Few-body physics in resonantly interacting ultracold quantum gases. *J. Phys. B* **51**, 043001 (2018).
24. Wolf, J. *et al.* State-to-state chemistry for three body recombination in an ultracold rubidium gas. *Science* **358**, 921–924 (2017).
25. Campbell, S. L. *et al.* A Fermi-degenerate three-dimensional optical lattice clock. *Science* **358**, 90–94 (2017).
26. Marti, G. E. *et al.* Imaging optical frequencies with $100\mu\text{Hz}$ precision and $1.1\mu\text{m}$ resolution. *Phys. Rev. Lett.* **120**, 103201 (2018).
27. Campbell, G. K. *et al.* Imaging the Mott insulator shells by using atomic clock shifts. *Science* **313**, 649–652 (2006).
28. Kato, S. *et al.* Laser spectroscopic probing of coexisting superfluid and insulating states of an atomic Bose–Hubbard system. *Nat. Commun.* **7**, 11341 (2016).
29. Zang, X.-P., Yang, M., Ozaydin, F., Song, W. & Cao, Z.-L. Generating multiatom entangled W states via light-matter interface based fusion mechanism. *Sci. Rep.* **5**, 16245 (2015).
30. Martinez de Escobar, Y. N. *et al.* Two-photon photoassociative spectroscopy of ultracold ^{88}Sr . *Phys. Rev. A* **78**, 062708 (2008).
31. Burt, E. A. *et al.* Coherence, correlations, and collisions: what one learns about Bose–Einstein condensates from their decay. *Phys. Rev. Lett.* **79**, 337–340 (1997).
32. Söding, J. *et al.* Three-body decay of a rubidium Bose–Einstein condensate. *Appl. Phys. B* **69**, 257–261 (1999).
33. Wang, J., D’Incao, J. P., Wang, Y. & Greene, C. H. Universal three-body recombination via resonant d-wave interactions. *Phys. Rev. A* **86**, 062511 (2012).

Acknowledgements We acknowledge technical contributions from W. Milner, E. Oelker, J. Robinson, L. Sonderhouse and W. Zhang, and discussions with T. Bothwell, S. Bromley, C. Kennedy, D. Kedar, S. Kolkowitz, M. D. Lukin, A. Safavi-Naini and C. Sanner. This work was supported by NIST, DARPA, W911NF-16-1-0576 through ARO, AFOSR-MURI, AFOSR, NSF-1734006 and NASA. A.G. is supported by a postdoctoral fellowship from the Japan Society for the Promotion of Science and G.E.M. is supported by a postdoctoral fellowship from the National Research Council. J.P.D. acknowledges support from NSF Grant PHY-1607204.

Author contributions A.G., R.B.H., G.E.M., S.L.C. and J.Y. contributed to the experiments. M.A.P., P.S.J., J.P.D. and A.M.R. contributed to the development of the theoretical model. All authors discussed the results, contributed to the data analysis and worked together on the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0661-6>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to A.G. and J.Y.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

State preparation. At the end of a 10-s evaporation, a ten-spin-component Fermi gas is loaded into a cubic state-independent optical lattice in a two-stage ramp. The first, 300-ms ramp to about $5E_{\text{rec}}$ is used consistently for all the measurements shown. To prepare for $n=4$ and $n=5$ occupied sites, the second ramp to the final lattice depth is sped up from 200 ms to 50 ms to minimize three-body loss during the loading process. n -occupied sites are randomly filled with n different nuclear-spin components from the $\binom{10}{n}$ nuclear-spin configurations. The initial entropy per particle in the lattice is estimated to be $s/k_{\text{B}}=1.8$ (k_{B} , Boltzmann constant) from the T/T_{F} measured in the dipole trap before and after lattice loading. In the atomic limit, in which tunnelling is negligible, the maximum spin entropy for $n=1$ sites is $s_{\text{spin}}/k_{\text{B}}=\ln(10)=2.3$ for ten spin states. This leads to a lowered temperature in the lattice when entropy is transferred from the motional to the spin degree of freedom^{10,11}.

To minimize a systematic shift due to the inhomogeneity of the trap depth across the cloud, we prepare a $10\text{ }\mu\text{m}\times 10\text{ }\mu\text{m}\times 2\text{ }\mu\text{m}$ sample of sites with the desired occupation by optimizing the final evaporation point. For each image, we measure the spectroscopic response in only the $4\text{ }\mu\text{m}\times 4\text{ }\mu\text{m}$ region at the centre of the lattice. As the on-site frequency shifts for large occupations increase, the line shapes become asymmetric owing to the residual inhomogeneity of the trap depth. To determine the peak frequencies, we fit each spectrum with an asymmetric Lorentzian, as shown in Fig. 2a. The trap depth in the central region of the lattice is calibrated by motional sideband spectroscopy of an $n=1$ sample, using the same procedure.

Data availability

The data that support the findings of this study are available within the paper.

Three-dimensional collective charge excitations in electron-doped copper oxide superconductors

M. Hepting¹, L. Chaix^{1,9}, E. W. Huang^{1,2}, R. Fumagalli³, Y. Y. Peng^{3,10}, B. Moritz¹, K. Kummer⁴, N. B. Brookes⁴, W. C. Lee⁵, M. Hashimoto⁶, T. Sarkar⁷, J.-F. He^{1,11}, C. R. Rotundu¹, Y. S. Lee¹, R. L. Greene⁷, L. Braicovich^{3,4}, G. Ghiringhelli^{3,8}, Z. X. Shen^{1*}, T. P. Devereaux^{1*} & W. S. Lee^{1*}

High-temperature copper oxide superconductors consist of stacked CuO₂ planes, with electronic band structures and magnetic excitations that are primarily two-dimensional^{1,2}, but with superconducting coherence that is three-dimensional. This dichotomy highlights the importance of out-of-plane charge dynamics, which has been found to be incoherent in the normal state^{3,4} within the limited range of momenta accessible by optics. Here we use resonant inelastic X-ray scattering to explore the charge dynamics across all three dimensions of the Brillouin zone. Polarization analysis of recently discovered collective excitations (modes) in electron-doped copper oxides^{5–7} reveals their charge origin, that is, without mixing with magnetic components^{5–7}. The excitations disperse along both the in-plane and out-of-plane directions, revealing its three-dimensional nature. The periodicity of the out-of-plane dispersion corresponds to the distance between neighbouring CuO₂ planes rather than to the crystallographic *c*-axis lattice constant, suggesting that the interplane Coulomb interaction is responsible for the coherent out-of-plane charge dynamics. The observed properties are hallmarks of the long-sought ‘acoustic plasmon’, which is a branch of distinct charge collective modes predicted for layered systems^{8–12} and argued to play a substantial part in mediating high-temperature superconductivity^{10–12}.

The charge dynamics of systems with periodically stacked quasi-two-dimensional (2D) conducting planes are strongly affected in the presence of poorly screened interplane Coulomb interactions. In a simple layered electron gas with conducting planes separated by dielectric spacers^{8,9}, the dispersion of plasmons (that is, the collective dynamical charge modes) changes from optical-like to acoustic-like as a function of out-of-plane momenta q_z (Fig. 1a) (such plasmons are referred to as ‘acoustic plasmons’ hereafter), a behaviour distinct from that in either pure 2D or isotropic 3D systems. For superconducting copper oxides, similar charge dynamics have been postulated because they consist of conducting CuO₂ planes stacked along the *c*-axis with poor out-of-plane Coulomb screening^{10–12}. Although plasmons have been observed in various spectroscopic studies at the Brillouin zone centre^{4,13,14} and also by transmission electron energy loss spectroscopy (EELS), which typically explores in-plane energy-momenta dispersions at $q_z=0$ (ref. ¹⁵), there is no information on its possible q_z dependence. Experimental evidence of this previously undetected component and its characterization in energy and momentum can shed new light on long-standing hypotheses that connect out-of-plane charge dynamics to superconductivity. For instance, it has been proposed that 20% of the observed value of the high superconducting transition temperature T_c of the copper oxides can be attributed to the presence of acoustic plasmons^{10–12}, where the large amount of energy stored in the interplane Coulomb interactions is related to the

substantial energy savings associated with the high superconducting transition temperature^{16,17}.

In this Letter, we focus our attention on the enigmatic ‘zone centre’ excitation previously discovered by Cu L-edge resonant inelastic X-ray scattering (RIXS) in the electron-doped copper oxide Nd_{2-x}Ce_xCuO₄ (NCCO) ($x=0.15$)^{5,6} and Sr_{1-x}La_xCuO₂ (ref. ⁷). In a different family of electron-doped copper oxides La_{2-x}Ce_xCuO₄ (LCCO) ($x=0.175$), we resolve spectral features that are similar to those of NCCO at representative in-plane momentum transfers q_{\parallel} (red-shaded peak in Fig. 1b, c), suggesting the universality of this collective mode in electron-doped copper oxides. Speculation about its origin has included intra-band transitions⁵, collective modes of a quantum phase⁶, and plasmons¹⁸. Although the mode has been suspected to be of charge character, a definitive assessment has not been possible owing to the inability to distinguish between charge and magnetic excitations in previous measurements^{19,20}.

We first identify the character of this excitation by determining the associated magnetic and charge contributions to the RIXS spectra. This can be uniquely achieved by resolving the polarization of both the incident and scattered photons¹⁹. Namely, magnetic excitations flip spins and necessarily change the angular momentum of the photons in the scattering process, that is, contribute to the crossed-polarization channel ($\sigma\pi$ or $\pi\sigma$). Conversely, charge excitations preserve the angular momentum of the photon and contribute to the parallel polarization channel ($\sigma\sigma$ or $\pi\pi$). Figure 1d, e shows polarization-resolved RIXS spectra for two different in-plane momenta. At $q_{\parallel}=(0.045, 0)$ the features of the zone centre excitation are fully suppressed for crossed polarizations ($\sigma\pi$) and the spectrum contains only the parallel polarization ($\sigma\sigma$) contribution. For larger momentum transfer $q_{\parallel}=(0.095, 0)$ the mode disperses towards higher energy (about 0.8 eV) (Fig. 1e) and the well-studied paramagnon excitation^{5–7,20} emerges on a lower energy scale (about 0.3 eV). As expected, the paramagnon yields spectral weight in both polarizations owing to the mixture of single spin-flip excitation, double spin-flip and incoherent particle–hole charge excitations, whose spectral weight increases with increasing doping concentration²¹. Importantly, the zone centre excitations, which are separated in energy from the paramagnons, still appear only for parallel polarization geometries. Thus, we first conclude that the zone centre excitations are a branch of pure charge modes.

A second insight can be obtained from a comprehensive mapping of the energy–momentum dispersion in all three dimensions of reciprocal space, in contrast to previous RIXS experiments that explored the projected in-plane momentum without focusing on the q_z dependence^{6–8}. Figure 2a, b shows the RIXS intensity maps as a function of momentum transfer along the hh - and h -directions (that is, along $(0, 0, l)-(h, h, l)$ and $(0, 0, l)-(h, 0, l)$, respectively) at $l=1$ and $l=1.65$. We denote

¹Stanford Institute for Materials and Energy Sciences, SLAC National Accelerator Laboratory and Stanford University, Menlo Park, CA, USA. ²Department of Physics, Stanford University, Stanford, CA, USA. ³Dipartimento di Fisica, Politecnico di Milano, Milan, Italy. ⁴European Synchrotron Radiation Facility (ESRF), Grenoble, France. ⁵Department of Physics, Binghamton University, Binghamton, NY, USA. ⁶Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Menlo Park, CA, USA. ⁷Department of Physics, Center for Nanophysics and Advanced Materials, University of Maryland, College Park, MD, USA. ⁸CNR-SPIN, Politecnico di Milano, Milan, Italy. ⁹Present address: Université Grenoble Alpes, CNRS, Institut Néel, Grenoble, France.

¹⁰Present address: Department of Physics and Seitz Materials Research Lab, University of Illinois, Urbana, IL, USA. ¹¹Present address: Department of Physics, University of Science and Technology of China, Hefei, China. *e-mail: zxshen@stanford.edu; tpd@stanford.edu; leews@stanford.edu

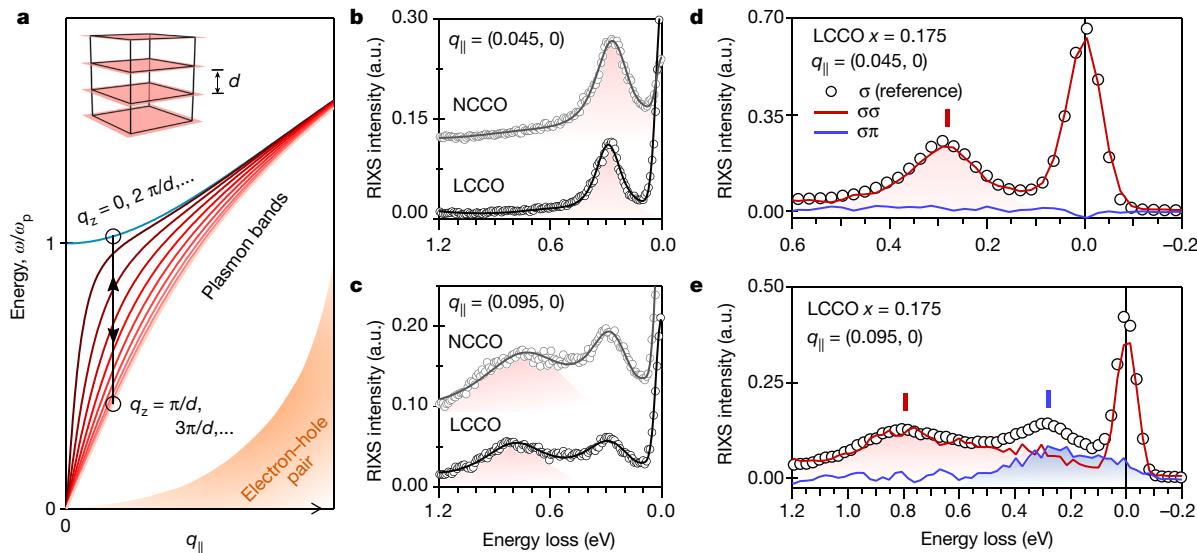


Fig. 1 | Plasmons in a layered electron gas and dispersive charge excitations in electron-doped copper oxides. **a**, Plasmon dispersion in a layered electron gas as a function of in- and out-of-plane momentum transfer q_{\parallel} and q_z , respectively. Different branches correspond to specific out-of-plane momentum transfers q_z and are a result of the interplane Coulomb interaction between the periodically stacked planes with distance d (see inset). Their spectrum varies from a single optical branch (light blue line) with the characteristic plasma frequency ω_p for $q_z=0$, $2\pi/d, \dots$ to a range of acoustic branches (light red lines) with the lowest-energy branch for $q_z=\pi/d, 3\pi/d, \dots$. The electron–hole pair excitation continuum is illustrated by the orange shaded area. **b, c**, RIXS spectra of NCCO ($x=0.15$) and LCCO ($x=0.175$) at in-plane momentum transfers $q_{\parallel}=(0.045, 0)$ and $(0.095, 0)$ for incident photon energies tuned

to the Cu L₃-edge at temperature $T \approx 20$ K. The spectral peak assigned to the dispersive zone centre excitation is highlighted by the red-shaded peak profile. The additional peak at about 0.3 eV in **c** is identified as the paramagnon, justified by the polarization-resolved RIXS spectra shown in **e**. The NCCO and LCCO spectra are offset in the vertical direction for clarity. **d, e**, Polarization-resolved RIXS spectra of LCCO at $q_{\parallel}=(0.045, 0)$ and $(0.095, 0)$. Charge excitations are detected in the parallel polarization channel ($\sigma\sigma$, red line) while magnetic excitations are detected in the spin-flip crossed-polarization channel ($\sigma\pi$, blue line). The reference spectrum (open symbols) is taken with σ polarized incident photons in the absence of polarization analysis, corresponding thus to the sum of $\sigma\sigma$ and $\sigma\pi$. Red and blue markers indicate the relevant spectral weight maximum of the $\sigma\sigma$ and $\sigma\pi$ channel, respectively. a.u., arbitrary units.

momentum transfer h, k, l in reciprocal lattice units ($2\pi/a, 2\pi/b, 2\pi/c$), where $a, b=4.01$ Å and $c=12.4$ Å are the lattice constants of LCCO with $x=0.175$. For both l , the excitations exhibit an almost linear dispersion emanating away from $(0, 0, l)$ along both the hh - and h directions. Surprisingly, while the paramagnons do not disperse appreciably with different l values (as expected for the quasi-2D magnetic structure of copper oxides), the dispersion of the zone centre excitation becomes steeper and increasingly separated from the paramagnon branch at larger l . This behaviour is highlighted in Fig. 2c. The raw data itself also clearly shows such an l -dependence: the spectral peak shifts to higher energy and becomes broader at the higher l -value for a given in-plane momentum (Fig. 2d, e, Extended Data Figs. 1, 2).

To further investigate the out-of-plane dependence, Fig. 3a displays the energy–momentum dispersion as a function of l at fixed in-plane momentum transfer $(0.025, 0)$ near the in-plane zone centre. As anticipated, the dispersion is symmetric around the out-of-plane zone centre $l=1$ as it is a high symmetry point in reciprocal space. Remarkably, the zone centre excitation continues dispersing towards high energy with further increasing l , insensitive to the next high symmetry point at $l=1.5$. In fact, the energy scale of the excitation continues to increase even at $l=1.8$, the highest out-of-plane momentum transfer that was accessible in our experiment. We also observe the same peculiar behaviour in our momentum-resolved RIXS measurements on NCCO (see Extended Data Fig. 3), suggesting a universal origin of the branch of charge modes from the three-dimensional (3D) nature of the Coulomb interaction in an otherwise layered quasi-2D material.

These results can be rationalized by doubling the out-of-plane Brillouin zone size, implying that the crystallographic unit cell with lattice constant c does not set the periodicity of the zone centre excitation, but rather $d=c/2$: the nearest-neighbour CuO₂ plane spacing (Fig. 3b). Hence, a new index l^* in units of $2\pi/d$ appropriately describes the Brillouin zone that is ‘felt’ by the zone centre excitation and establishes the proper periodicity of the dispersion. An obvious mechanism

that could induce such a Brillouin zone ‘reconstruction’ in a quasi-2D system is the interplanar Coulomb interaction.

These striking features are reminiscent of acoustic plasmon bands theoretically proposed in the aforementioned layered electron gas model (Fig. 1a), which provides a qualitative fit to the observed zone centre excitation (see Methods and Extended Data Fig. 4 for the fits). To demonstrate the behaviour of collective charge dynamics beyond this weakly correlated layered electron gas model, we perform determinant quantum Monte Carlo (DQMC) calculations for a 2D three-band Hubbard model with an electron doping of $x=0.18$ and incorporate three-dimensionality through interplane Coulomb interactions using a random-phase-approximation-like formalism (see Methods). Therewith, we obtain the collective charge response via the loss function $-\text{Im}(1/\varepsilon)$, where ε is the dielectric function, versus the out-of-plane momentum transfer l^* . As shown in Fig. 3c, the calculated dispersions along the l^* -direction at the in-plane momenta accessible in our 16×4 cluster exhibit the same qualitative behaviours observed in both LCCO (Fig. 3a) and NCCO (Extended Data Fig. 3). We note that recent calculations using different methods and models have also demonstrated l^* -dependent plasmon bands^{18,22}. Although RIXS is not simply proportional to $-\text{Im}(1/\varepsilon)$ owing to the resonant process^{19,23}, it nevertheless can contain critical information on the loss function, such as the dispersion of excitations. Thus, the agreement between our data and theory lends strong support to the idea of attributing the zone centre mode to plasmon excitations, with momentum-resolved RIXS providing access to its acoustic bands.

Interestingly, the data shown in Figs. 2, 3a indicate that the plasmon peak broadens when approaching the equivalent zone centre $(0, 0, l^*=1)$ along the l^* -direction (see also Extended Data Figs. 1c, 3c). In fact, at momentum transfer $(0.025, 0, l^*=0.925)$, that is, close to the equivalent zone centre $(0, 0, l^*=1)$, the plasmon linewidth of approximately 0.5 eV (Fig. 3a) is similar to previous transmission EELS reports¹⁵ and optical conductivity measurements at $(0, 0, 0)$ ¹⁴. The increasing incoherence of the mode near the zone centre is consistent

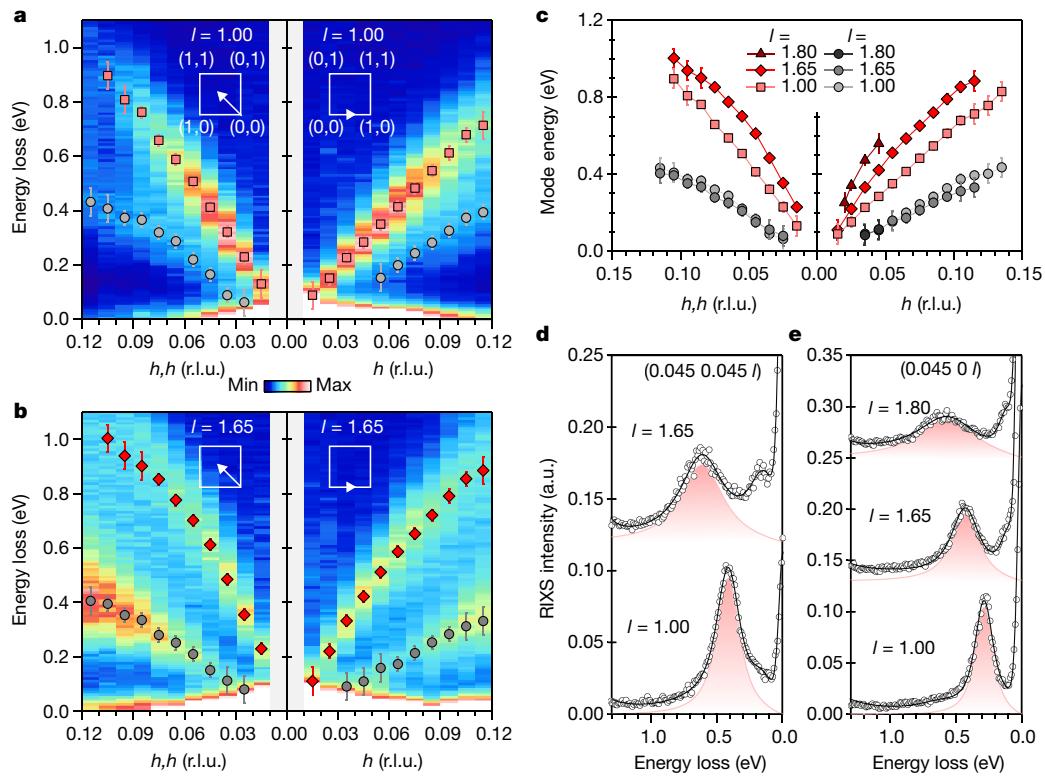


Fig. 2 | Three-dimensionality of the zone centre excitations.

a, b, RIXS intensity maps of LCCO ($x=0.175$) for momentum transfer along the hh - and h -directions at $l=1$ and $l=1.65$. Red and grey symbols indicate least-squares-fit peak positions of the zone centre excitation and the paramagnon, respectively (see Extended Data Figs. 1 and 2). Error bars are estimated from the uncertainty in energy-loss reference-point determination (± 0.01 eV) together with the standard deviation of the fits. The insets indicate the probe direction in reciprocal space. **c**, Summary of the energy dispersion of the zone centre excitation (red symbols) and the paramagnon

with incoherent charge dynamics inferred from c -axis optical conductivity^{3,4}. However, such linewidth evolution appears to be different from the Landau quasi-particle picture in which the plasmon peak should be sharpest at the zone centre where the particle-hole (Landau) continuum is minimal and well separated from the plasmon. Other mechanisms, such as the presence of non-Fermi-liquid-producing interactions, polar interlayer electron-phonon coupling²⁴ and Umklapp scattering²⁵, may reconcile these observations.

We now investigate the acoustic plasmon bands in LCCO as a function of the carrier density, which is nominally the Ce concentration x . We have also verified the systematics of doping concentration among different samples using an internal spectral reference—the dd excitations of RIXS spectra (Methods and Extended Data Fig. 5). As shown in Fig. 4a, b, the plasmon bands exhibit a detectable doping dependence. In a naive picture of the free electron model, the plasmon energy is expected to increase proportionally to $\sqrt{x/m^*}$, where m^* is the effective electron mass. Consistent with such an expectation the mode energies of $x=0.11$ to about 0.15 increase linearly with \sqrt{x} (Fig. 4c), further substantiating the attribution of the zone centre excitation to a plasmon. However, for higher dopings, the rate of increase slows down and appears to hit a plateau for $x=0.17$ and 0.18. This observation suggests a possible variation in the band dispersion or Fermi surface at approximately $x \approx 0.15$. Recent Hall-effect measurements on LCCO have indicated Fermi-surface reconstruction due to antiferromagnetic correlations ending at around $x \approx 0.14$ (ref. ²⁶), corroborating our observation.

We note that our observation appears to be distinct from a recent high-resolution EELS measurement, which reported featureless in-plane charge excitations in a hole-doped copper oxide²⁷. This

(grey symbols) for different l -values, that is, for different momentum transfers along the c -axis. The energy dispersion of the paramagnon is independent of l within the experimental error. The lines connecting markers serve as a guide for the eye. **d, e**, Representative raw RIXS spectra (open symbols) for momentum transfer along the hh - and h -directions at different l -values, together with the anti-symmetrized Lorentzian fit profiles of the zone centre excitation (red shades) and the sum of all contributions fitted to the spectra (solid black lines, see also Extended Data Figs. 1 and 2). Spectra are offset in the vertical direction for clarity. r.l.u., reciprocal lattice unit.

suggests a distinct behaviour of the charge degrees of freedom between electron- and hole-doped copper oxides. However, given the similarities of the layered 2D CuO_2 plane structure between electron- and hole-doped materials, we should expect a similar three-dimensionality in the charge dynamics of the hole-doped copper oxides. Although previous RIXS studies on hole-doped compounds primarily focused on magnetic, orbital and other high-energy excitations²⁰, few investigated the region of energy-momentum space necessary to identify and characterize the acoustic plasmon²⁸. More detailed RIXS measurements on hole-doped compounds, in both single and multi-layer systems with higher T_c , will be able to clarify these issues.

Our observation of 3D plasmon modes indicates that the copper oxides, at the very least, should be modelled as layered 2D systems when describing their charge dynamics, a fact that is often overlooked. This change of perspective has important implications. First, in a 2D doped Mott insulator, the in-plane charge fluctuations are strongly suppressed by the Coulomb interaction. Thus the spin dynamics become the most prominent low-energy excitations, thought to be most relevant to high- T_c superconductivity²⁹. Our results challenge this view by demonstrating that the low energy charge fluctuations can be quite active owing to the layered structure of the copper oxide. Early theories suggested that acoustic plasmons may be able to mediate pairing¹⁰, or perhaps more importantly, enhance T_c for Cooper pairs bound by other interactions¹². Second, the Coulomb energy stored between CuO_2 planes can be important and may play a part in the energy savings associated with the superconducting transition^{16,17}. Third, the energy of the plasmon extrapolates to approximately zero at the projected zone centre. This implies a negligible single electron hopping between adjacent CuO_2 planes, leaving the Coulomb interaction as the sole

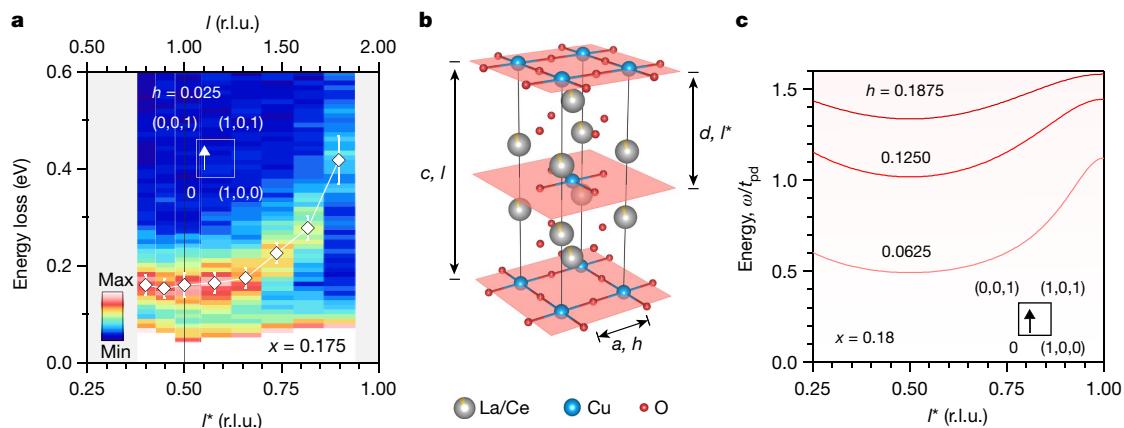


Fig. 3 | Out-of-plane plasmon dispersion. **a**, RIXS intensity map of LCCO ($x = 0.175$) for momentum transfer along the out-of-plane direction at $h = 0.025$. The out-of-plane momentum is indicated by the indices l (top scale, units of $2\pi/c$) corresponding to the crystallographic c -axis, and l^* (bottom scale, units of $2\pi/d$) corresponding to the CuO_2 plane spacing. White symbols indicate fitted peak positions of the zone centre excitation. The black vertical line highlights the high symmetry point at $l^* = 0.5$ (that is, $l = 1.0$). Error bars are estimated from the uncertainty in energy-loss reference-point determination (± 0.01 eV)

source of interplanar coupling. Such restriction of the charge to the 2D planes may enhance the effects of quantum confinement at the heart of topological theories for superconductivity in copper oxides^{30,31}. However, important questions remain about the impact of the interplanar Coulomb interaction on the electronic structure, the pseudogap and charge- and spin-density-wave orders. In a broader context, our result fits within the general framework that the Coulomb interaction

together with the standard deviation of the fits. **b**, Crystal structure of LCCO with the crystallographic unit cell indicated by black lines and the CuO_2 planes in red. **c**, Calculation of the charge dynamics for planes of a stacked three-band Hubbard model (see Methods) with electron doping $x = 0.18$. The peak frequency of the loss function $-\text{Im}(1/\varepsilon)$ is plotted versus the out-of-plane momentum transfer l^* at $h = 0.1875$, 0.1250 and 0.0625. The energy is expressed in units of t_{pd} , the hopping integral between the oxygen $2p$ and Cu $3d$ orbitals in the three-band Hubbard model.

affects charge dynamics in nanoscale heterostructures of 2D quantum materials, including transitional metal chalcogenides and graphene. Manipulating the Coulomb interaction could enable fine-tuning of desired properties of artificial quantum materials, such as plasmonics in nanostructures, with a range of applications from sensors to photonic and electronic devices for communications (for example, see the Plasmonics Focus Issue introduced in ref. ³²).

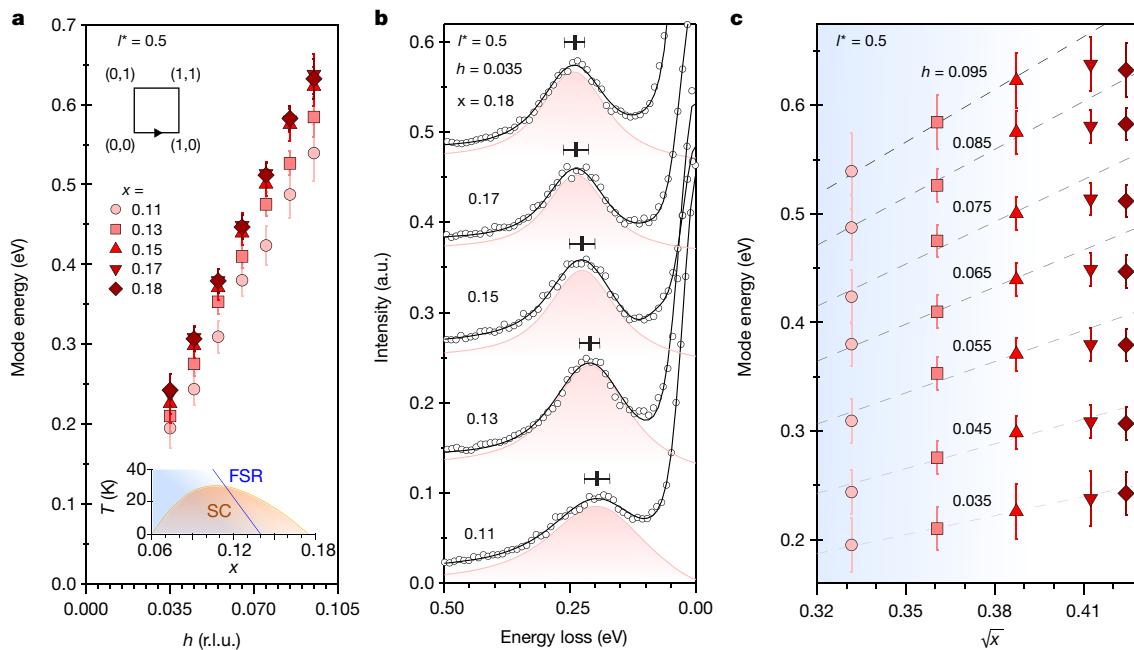


Fig. 4 | Doping dependence of the plasmon. **a**, Energy dispersion of the LCCO zone centre mode for momentum transfer along the h -direction at $l^* = 0.5$. Red symbols are the fitted peak positions for Ce doping concentrations $x = 0.11, 0.13, 0.15, 0.17$ and 0.18 . The bottom inset shows the temperature versus doping phase diagram of LCCO thin films adapted from ref. ²⁹ including a superconducting dome (orange shading, SC) and a small Fermi surface region (blue shading) undergoing a crossover to a reconstructed Fermi surface (blue line, FSR) that ends at $x \approx 0.14$. **b**, Representative raw RIXS spectra (open symbols) for momentum transfer $h = 0.035$ and $l^* = 0.5$ for different dopings. The anti-symmetrized Lorentzian fit profiles are shaded in red with vertical black markers

indicating the peak positions and the horizontal bars indicating the error bars. Spectra are offset in the vertical direction for clarity. **c**, Mode energies versus the square root of x for in-plane momenta from $h = 0.035$ to 0.095 at $l^* = 0.5$. The symbol style refers to the dopings as indicated in **a**. Dashed lines are linear fits of the mode energy versus \sqrt{x} for dopings $x = 0.11$ to 0.15 , constrained by assuming that the mode energy is 0 at $x = 0$. We note that the rate of increase deviates from linear behaviour and forms a plateau for higher doping, as mentioned in the text. Error bars in this figure are estimated from the uncertainty in energy-loss reference-point determination (± 0.01 eV) together with the standard deviation of the fits.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0648-3>.

Received: 2 May 2018; Accepted: 22 August 2018;

Published online 31 October 2018.

1. Damascelli, A., Hussain, Z. & Shen, Z. X. Angle-resolved photoemission studies of the cuprate superconductors. *Rev. Mod. Phys.* **75**, 473 (2003).
2. Fujita, M. et al. Progress in neutron scattering studies of spin excitations in high- T_c cuprates. *J. Phys. Soc. Jpn.* **81**, 011007 (2012).
3. Tajima, S. Optical studies of high-temperature superconducting cuprates. *Rep. Prog. Phys.* **79**, 094001 (2016).
4. Singley, E. J., Basov, D. N., Kurahashi, K., Uefuji, T. & Yamada, K. Electron dynamics in $\text{Nd}_{1.85}\text{Ce}_{0.15}\text{CuO}_{4+\delta}$: evidence for the pseudogap state and unconventional c-axis response. *Phys. Rev. B* **64**, 224503 (2001).
5. Ishii, K. et al. High-energy spin and charge excitations in electron-doped copper oxide superconductors. *Nat. Commun.* **5**, 3714 (2014).
6. Lee, W. S. et al. Asymmetry of collective excitations in electron- and hole-doped cuprate superconductors. *Nat. Phys.* **10**, 883 (2014).
7. Dellea, G. et al. Spin and charge excitations in artificial artificial hole- and electron-doped infinite layer cuprate superconductors. *Phys. Rev. B* **96**, 115117 (2017).
8. Grecu, D. Plasmon frequency of the electron gas in layered structures. *Phys. Rev. B* **8**, 1958 (1973).
9. Fetter, A. L. Electrodynamics of a layered electron gas II. Periodic array. *Ann. Phys.* **88**, 1 (1974).
10. Kresin, V. Z. & Morawitz, H. Layer plasmons and high- T_c superconductivity. *Phys. Rev. B* **37**, 7854 (1988).
11. Ishii, Y. & Ruvalds, J. Acoustic plasmons and cuprate superconductivity. *Phys. Rev. B* **48**, 3455 (1993).
12. Bill, A., Morawitz, H. & Kresin, V. Z. Electronic collective modes and superconductivity in layered conductors. *Phys. Rev. B* **68**, 144519 (2003).
13. Bozovic, I. Plasmons in cuprate superconductors. *Phys. Rev. B* **42**, 1969 (1990).
14. Levallois, J. et al. Temperature-dependent ellipsometry measurements of partial coulomb energy in superconducting cuprates. *Phys. Rev. X* **6**, 031027 (2016).
15. Fink, J., Knupfer, M., Atzkern, S. & Golden, M. Electronic correlation in solids, studies using electron energy-loss spectroscopy. *J. Elec. Spectrosc. Rel. Phenom.* **117/118**, 287–309 (2001).
16. Leggett, A. J. Where is the energy saved in cuprate superconductivity? *J. Phys. Chem. Solids* **59**, 1729 (1998).
17. Leggett, A. J. Cuprate superconductivity: dependence of T_c on the c-axis layering structure. *Phys. Rev. Lett.* **83**, 392 (1999).
18. Greco, A., Yamase, H. & Bejas, M. Plasmon excitations in layered high- T_c cuprates. *Phys. Rev. B* **94**, 075139 (2016).
19. Jia, C., Wohlfeld, K., Wang, Y., Moritz, B. & Devereaux, T. P. Using RIXS to uncover elementary charge and spin excitations. *Phys. Rev. X* **6**, 021020 (2016).
20. Ament, L. J. P., van Veenendaal, M., Devereaux, T. P., Hill, J. P. & van den Brink, J. Resonant inelastic X-ray scattering studies of elementary excitations. *Rev. Mod. Phys.* **83**, 705 (2011).
21. Huang, H. Y. et al. Raman and fluorescence characteristics of resonant inelastic X-ray scattering from doped superconducting cuprates. *Sci. Rep.* **6**, 19657 (2016).
22. Markiewicz, R. S., Hasan, M. Z. & Bansil, A. Acoustic plasmons and doping evolution of Mott physics in resonant inelastic X-ray scattering from cuprate superconductors. *Phys. Rev. B* **77**, 094518 (2008).
23. Kim, J. et al. Comparison of resonant inelastic X-ray scattering spectra and dielectric loss functions in copper oxides. *Phys. Rev. B* **79**, 094525 (2009).
24. Meevasana, W., Devereaux, T. P., Nagaosa, N., Shen, Z. X. & Zaanen, J. Calculation of overdamped c-axis charge dynamics and the coupling to polar phonons in cuprate superconductors. *Phys. Rev. B* **74**, 174524 (2006).
25. Lee, W. C. Superconductivity-induced changes in density-density correlation function enabled by Umklapp processes. *Phys. Rev. B* **91**, 224503 (2015).
26. Sarkar, T. et al. Fermi surface reconstruction and anomalous low-temperature resistivity in electron-doped $\text{La}_{1-x}\text{Ce}_x\text{CuO}_4$. *Phys. Rev. B* **96**, 155449 (2017).
27. Mitrano, M. et al. Anomalous density fluctuation in a strange metal. *Proc. Natl. Acad. Sci. USA* **115**, 5392 (2018).
28. Ishii, K. et al. Observation of momentum-dependent charge excitations in hole-doped cuprates using resonant inelastic X-ray scattering at the oxygen K edge. *Phys. Rev. B* **96**, 115148 (2017).
29. Scalapino, D. J. et al. A common thread: the pairing interaction for unconventional superconductors. *Rev. Mod. Phys.* **84**, 1383 (2012).
30. Clarke, D. G., Strong, S. P. & Anderson, P. W. Incoherence of single particle hopping between Luttinger liquids. *Phys. Rev. Lett.* **72**, 3218 (1994).
31. Sachdev, S. & Chowdhury, D. The novel metallic states of the cuprates: Fermi liquids with topological order, and strange metals. *Prog. Theor. Exp. Phys.* **2016**, 12C102 (2016).
32. Pile, D. Perspective on plasmonics. *Nat. Photon.* **6**, 714–715 (2012).

Acknowledgements This work is supported by the US Department of Energy (DOE), Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division, under contract DE-AC02-76SF00515. L.C. acknowledges support from the Department of Energy, SLAC Laboratory Directed Research and Development funder contract under DE-AC02-76SF00515. RIXS data were taken at beamline ID32 of the European Synchrotron Radiation Facility (ESRF, Grenoble, France) using the ERIKS spectrometer designed jointly by the ESRF and the Politecnico di Milano. G.G. and Y.Y.P. were supported by the ERC-P-ReXS project (2016-0790) of the Fondazione CARIPLO and Regione Lombardia, in Italy. R.L.G. and T.S. acknowledge support from NSF award DMR-1708334. Computational work was performed on the Sherlock cluster at Stanford University and on resources of the National Energy Research Scientific Computing Center, supported by the US DOE under contract number DE-AC02-05CH11231.

Reviewer information *Nature* thanks D. M. Casa, D. van der Marel and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions W.S.L., G.G., L.B., T.P.D. and Z.X.S. conceived the experiment. M. Hepting, W.S.L., L.C., R.F., Y.Y.P., G.G., M. Hashimoto, K.K. and N.B.B. conducted the experiment at ESRF. M. Hepting, L.C. and W.S.L. analysed the data. E.W.H., W.C.L., B.M. and T.P.D. performed the theoretical calculations. T.S., J.-F.H., C.R.R., Y.S.L. and R.L.G. synthesized and prepared samples for the experiments. M. Hepting, B.M. and W.S.L. wrote the manuscript with input from all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0648-3>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0648-3>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to Z.X.S. or T.P.D. or W.S.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

The c -axis-oriented LCCO thin films with Ce concentrations $x=0.11, 0.13, 0.15, 0.17, 0.175$ and 0.18 were fabricated on (100) SrTiO₃ substrates by pulsed laser deposition using a KrF excimer laser. The annealing process was optimized for each x . The superconducting T_c of the $x=0.11$ and 0.13 films is around 30 K and around 22 K, respectively. The $x=0.175$ and 0.18 films did not show a superconducting transition. The NCCO single crystal with Ce concentration $x=0.15$ was grown by the travelling-solvent floating-zone method in O₂ and annealed in Ar at 900 °C for 10 h. The T_c is about 26 K.

The RIXS measurements were performed at beamline ID32 of the ESRF using the high-resolution 'ERIXS' spectrometer. The scattering angle 2θ can be changed in a continuous way from 50° to 150°. The samples were mounted on the 6-axis in-vacuum Huber diffractometer/manipulator and cooled to around 20 K. The RIXS data were obtained with incident σ polarization (perpendicular to the scattering plane, high-throughput configuration). The incident photon energy was tuned to the maximum of the Cu L_3 absorption peak at about 931 eV. The energy resolution was $\Delta E \approx 60$ meV for the $x=0.11, 0.15, 0.17$ and 0.18 LCCO sample and $\Delta E \approx 68$ meV for the $x=0.13$ and 0.175 LCCO and the $x=0.15$ NCCO sample. The RIXS polarization-resolved measurements were conducted with a wider monochromator exit slit at a resolution of $\Delta E \approx 85$ meV in order to partly compensate the reduced efficiency of the polarimeter with respect to the normal configuration. For each transferred momentum a non-resonant silver paint or carbon tape spectrum provided the exact position of the elastic (zero energy loss) line. For the polarimetric RIXS measurements of Fig. 1d, e, σ -polarization incident on the sample was used and the graded multilayer served as analyser of the scattered photons, as explained in detail in ref.³³. We note that Cu L-edge RIXS is the so-called direct RIXS process, involving the resonant transition from the Cu $2p$ core level to the $3d_{x^2-y^2}$ orbital, which constitutes the electronic structure near the Fermi energy. The Cu L-edge RIXS is capable of probing a wide range of elementary excitations, including the orbital (that is, the dd -excitations), magnetic excitations, phonons and charged excitations²⁰.

In the theory calculations we consider the three-band Hubbard model with the following standard parameters in units of electronvolts: $U_{dd}=8.5$, $U_{pp}=4.1$, $t_{pd}=1.13$, $t_{pp}=0.49$, $\Delta_{pd}=3.24$ (ref.³⁴). Determinant quantum Monte Carlo (DQMC)³⁵ is used to solve the model on a fully periodic 16×4 cluster at a temperature of $T=0.125$ eV. For each doping, 512 independently seeded Markov chains with 50,000 measurements each are run. The charge susceptibilities obtained by DQMC are analytically continued to real frequency using the maximum entropy method with model functions determined by the first moments of the data³⁶. The inverse dielectric function plotted in Fig. 3c is obtained via $\frac{1}{\varepsilon(q, \omega)} = \frac{1}{1 + V_q \chi(q, \omega)}$,

where $\chi(q, \omega)$ is the real frequency charge susceptibility obtained from DQMC and the maximum entropy method. The long-range and 3D Coulomb interactions neglected in the three-band Hubbard model are captured by V_q . We use the layered electron gas form³⁷: $V_q = \frac{d}{2\varepsilon_\infty} \frac{\sinh(q_\parallel d)}{q_\parallel [\cosh(q_\parallel d) - \cos(q_z d)]}$ where d is the interplane spacing and q_\parallel and q_z are the in-plane and out-of-plane components of the momentum transfer, respectively. ε_∞ , the sole free parameter of our calculation, is adjusted to give a roughly 1-eV mode for $q_z=0$ and the smallest non-zero $q_\parallel=(0.0625, 0)$; its value is not varied with doping.

Fit of the plasmon dispersion in the layered electron gas model. We consider the energy-momentum dispersion of a plasmon mode in a layered electron gas model for momentum transfer q_\parallel along the h -direction at fixed out-of-plane momentum transfer values q_z (along the l^* -direction). Let d be the spacing between adjacent CuO₂ planes and $q_z d=\pi/l^*$ and ε_∞ be the high-frequency dielectric constant due to the screening by the core electrons. Following ref.³⁸, the Coulomb potential of a layered electron gas is:

$$V_q = \alpha^2 \frac{\sinh(q_\parallel d)}{[q_\parallel [\cosh(q_\parallel d) - \cos(q_z d)]}$$

with

$$\alpha = \sqrt{\frac{e^2 d}{2\varepsilon_0 \varepsilon_\infty}}$$

In an isotropic medium it is well-known that the 3D Coulomb potential is $e^2/\varepsilon_0 \varepsilon_\infty q^2$, while in a 2D plane the Coulomb potential is $e^2/2\varepsilon_0 \varepsilon_\infty q_\parallel$. These are the two limits of the above form of the layered electron gas, with V_q becoming $V_q = e^2/\varepsilon_0 \varepsilon_\infty q_\parallel^2$ in the approximation of long wavelengths ($q_z d \ll 1$ and $q_\parallel d \ll 1$), and $V_q = e^2/2\varepsilon_0 \varepsilon_\infty q_\parallel$ for short wavelengths ($q_z d \gg 1$, independent of q_z momentum).

Here, we use the full form from above and obtain the energy of the plasmon mode by:

$$\omega_p \approx q_\parallel \sqrt{V_q} \approx \alpha \sqrt{\frac{q_\parallel \sinh(q_\parallel d)}{\cosh(q_\parallel d) - \cos(q_z d)}}$$

With this expression we performed least-squares fits of the plasmon energy dispersion, simultaneously for the datasets of $l^*=0.5$, $l^*=0.825$ and $l^*=0.9$ with the global fitting parameter α , yielding $\alpha=2.06$ (Extended Data Fig. 4). We note that the layered electron gas model does not include strong correlations, which explains the deviations between the experimental data and the model fit.

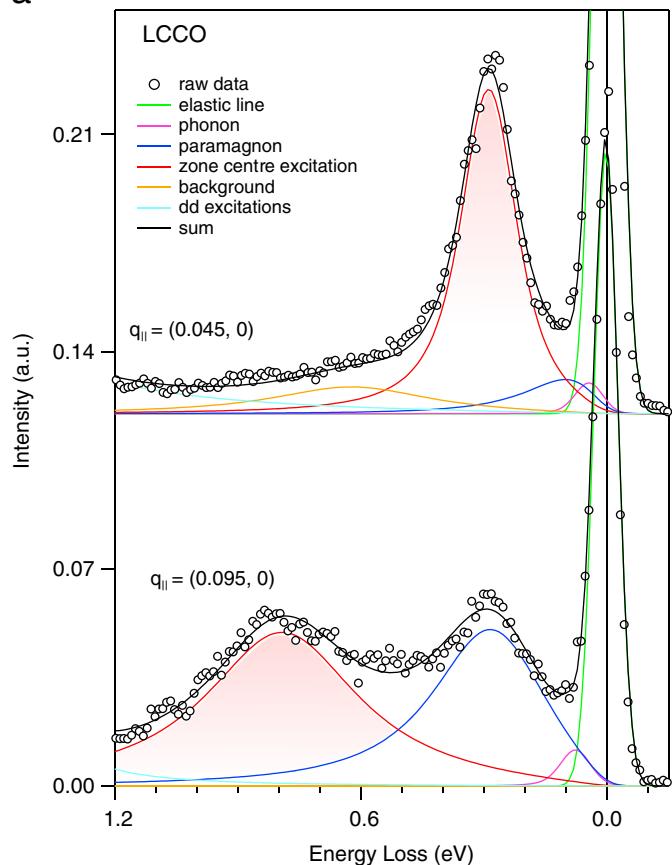
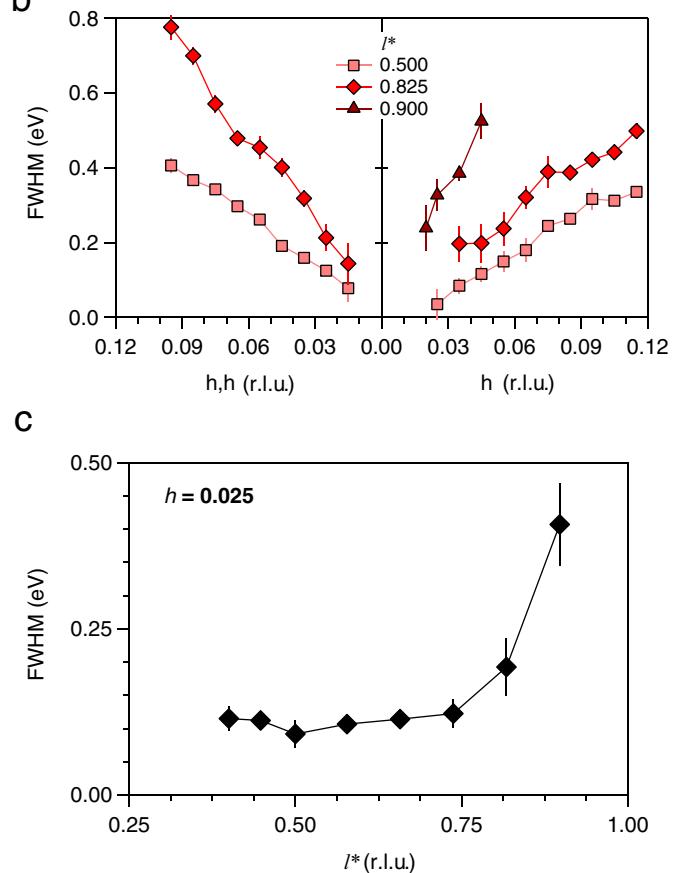
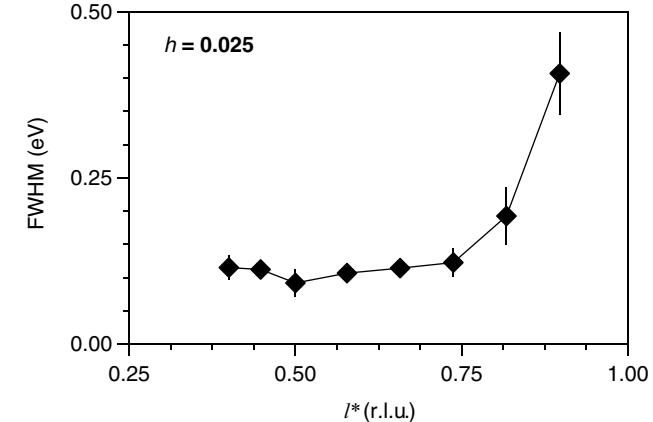
Verifying the systematics of electron doping. To confidently plot the doping dependence shown in Fig. 4, we verified the systematics of electron doping concentration using an internal reference of RIXS spectra—the energy positions of the dd excitations located in the energy range of 1.5–3 eV (Extended Data Fig. 5a). dd excitations are transitions within Cu $3d$ orbitals; thus, they are forbidden by optical dipole transitions, but allowed in the RIXS process³⁹. Because electron-doping primarily fills the Cu $3d_{x^2-y^2}$ orbital, the energy separation between the Fermi energy E_F and other occupied d -orbitals increases with increasing electron doping. As a consequence, the energy position of the dd excitations—which is essentially the energy separation between E_F and occupied d orbitals—shifts to higher energy loss in the RIXS spectra (Extended Data Fig. 5a and its inset). This shift should correlate with the nominal Ce doping concentration x , providing an independent verification of the electron doping among different samples. Indeed, as shown in Extended Data Fig. 5b, all our samples of $x=0.11, 0.13, 0.15, 0.17$ and 0.18 correlate well with x , except for the $x=0.175$ sample. This indicates that the actual electron doping concentration of the nominally $x=0.175$ sample might be affected by some change of oxygen composition or local inhomogeneity. Therefore we did not include the $x=0.175$ data in Fig. 4. We emphasize that the qualitative behaviours, that is, the out-of-plane dependence shown in Figs. 1–3, are the same for all doping concentrations of LCCO and also for the NCCO (Extended Data Fig. 3) that were studied in this work.

Code availability. Source code for the DQMC simulations is available at <https://github.com/cmendl/hubbard-dqmc>. Sample input files are included for the three-band Hubbard model studied in this work.

Data availability

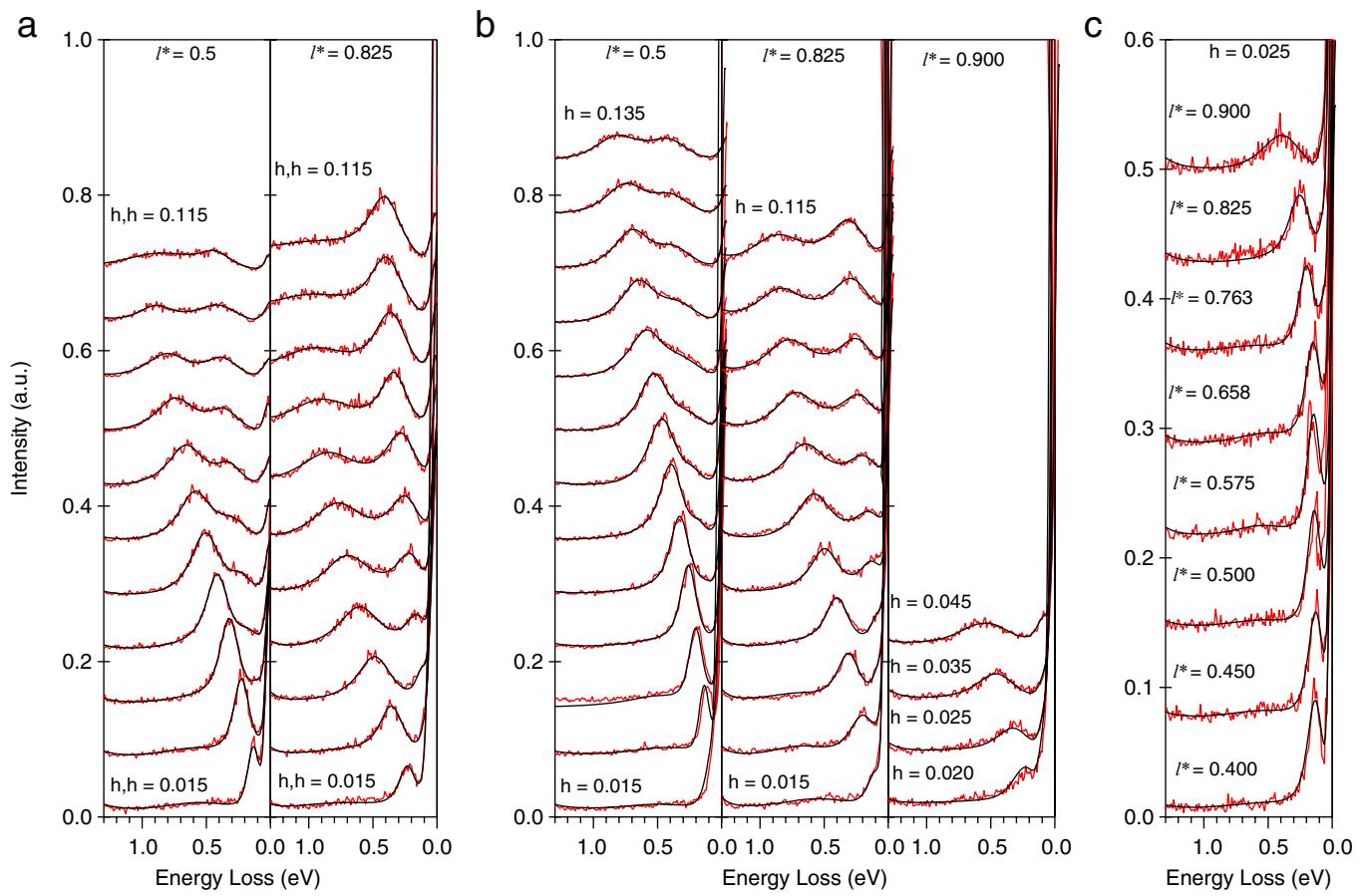
Raw data are included for Figs. 1b–e, 2, 3a, 4, and Extended Data Figs. 1–4. The data that support the plots within this paper and other findings of this study are available from the corresponding authors upon reasonable request.

33. Braicovich, L. et al. The simultaneous measurement of energy and linear polarization of the scattered radiation in resonant inelastic soft X-ray scattering. *Rev. Sci. Instrum.* **85**, 115104 (2014).
34. Kung, Y. F. et al. Characterizing the three-orbital Hubbard model with determinant quantum Monte Carlo. *Phys. Rev. B* **93**, 155166 (2016).
35. Blankenbecler, R., Scalapino, D. J. & Sugar, R. L. Monte Carlo calculations of coupled boson-fermion systems. I. *Phys. Rev. D* **24**, 2278 (1981).
36. Jarrell, M. & Gubernatis, J. E. Bayesian inference and the analytic continuation of imaginary-time Monte Carlo data. *Phys. Rep.* **269**, 133 (1996).
37. Fetter, A. L. Electrodynamics of a layered electron gas. I. Single layer. *Ann. Phys.* **81**, 367 (1973).
38. Turlakov, M. & Leggett, A. J. Sum rule analysis of umklapp processes and Coulomb energy: application to cuprate superconductivity. *Phys. Rev. B* **67**, 094517 (2003).
39. Moretti Sala, M. et al. Energy and symmetry of dd excitations in undoped layered cuprates measured by Cu L_3 resonant inelastic X-ray scattering. *New J. Phys.* **13**, 043026 (2011).
40. Le Tacon, M. et al. Intense paramagnon excitations in a large family of high-temperature superconductors. *Nat. Phys.* **7**, 725 (2011).

a**b****c**

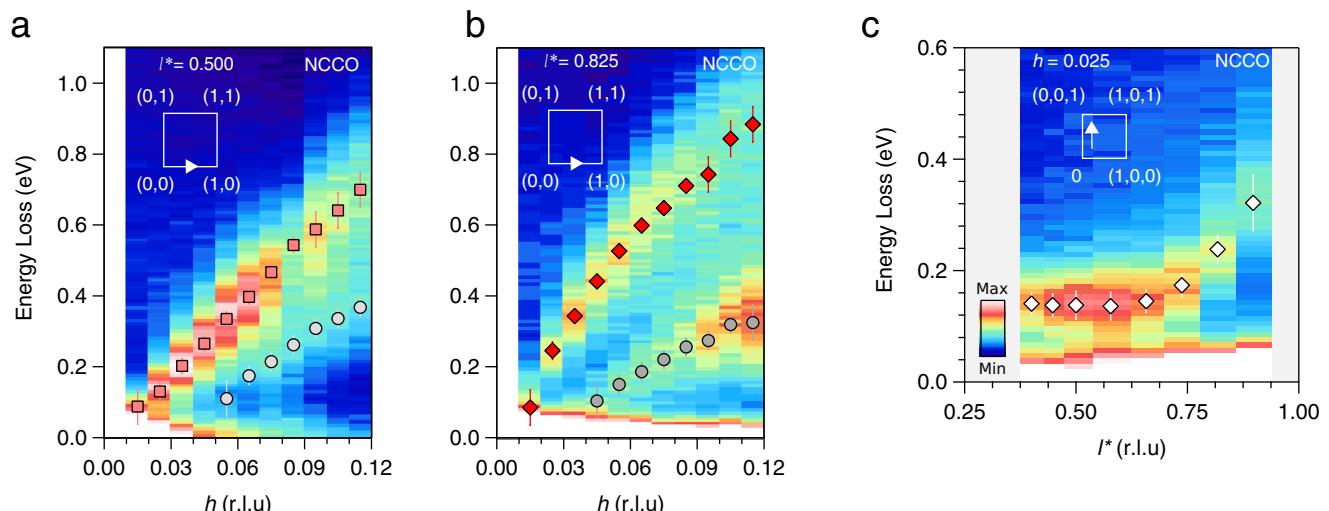
Extended Data Fig. 1 | Fits of the RIXS spectra. **a**, Fits of LCCO ($x=0.175$) RIXS spectra at in-plane momentum transfer positions $q_{\parallel}=(0.045, 0)$ and $(0.095, 0)$, representative of all fits performed in the scope of this work. The model uses a Gaussian for the elastic peak (green) and anti-symmetrized Lorentzians for all other contributions in the spectrum, convoluted with the energy resolution (here $\Delta E=68$ meV) via Gaussian convolution. The anti-symmetrized Lorentzian is used to ensure zero mode intensity at zero energy loss, as explained in the supplementary information of ref. ⁴⁰. The peak profiles of the zone centre

excitation (plasmon) are shaded in red. **b**, Full-width at half-maximum (FWHM) of the zone centre excitation (plasmon) as extracted from the fits for momentum transfer along the hh - and h -directions at $l^*=0.5$, $l^*=0.825$ and $l^*=0.9$, corresponding to the fitted peak positions shown in Fig. 2c. Error bars are the standard deviation of the fits. **c**, FWHM of the zone centre excitation (plasmon) as extracted from the fits for momentum transfer along the out-of-plane direction at $h=0.025$. The panel corresponds to the fitted peak positions shown in Fig. 3a.



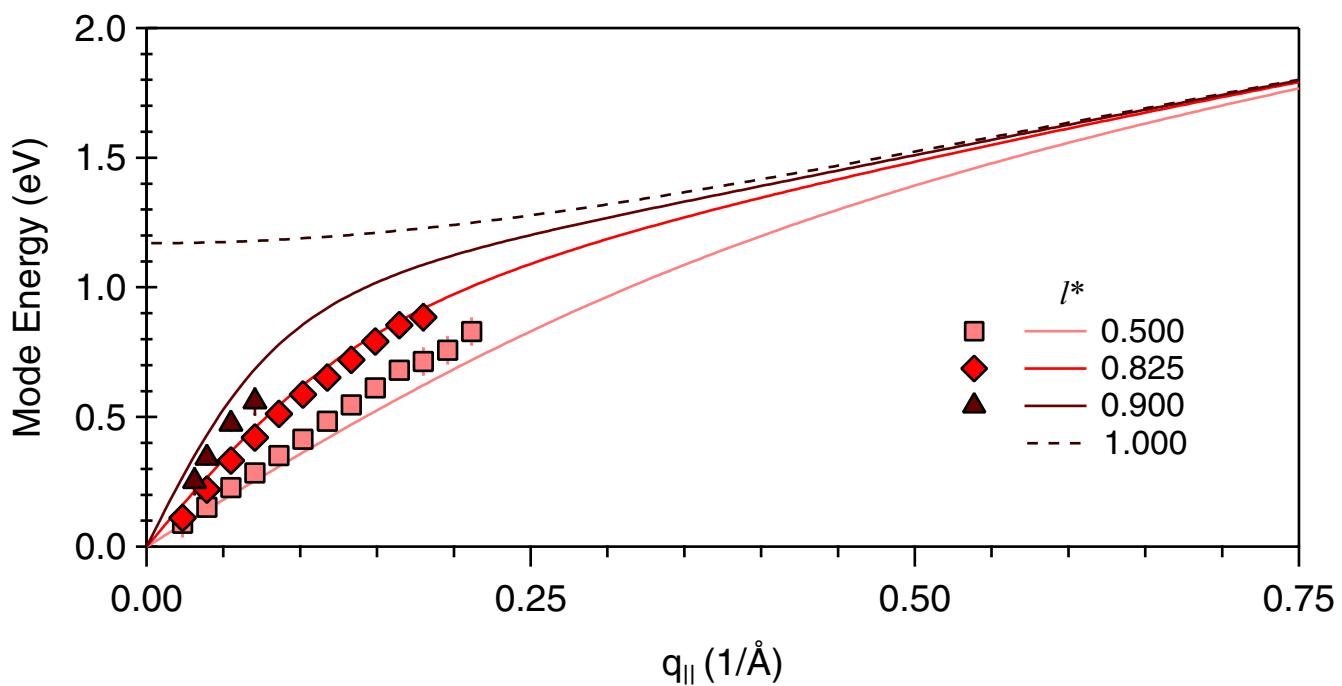
Extended Data Fig. 2 | Raw data and fits of the RIXS spectra. **a, b,** Raw RIXS spectra (red) of LCCO ($x = 0.175$) together with the fits (solid black lines) for momentum transfer along the hh -direction (**a**) and h -direction (**b**)

at different l^* . The spectra are offset in the vertical direction for clarity. **c,** Raw RIXS spectra together with the fits for momentum transfer along the l^* -direction at $h = 0.025$.



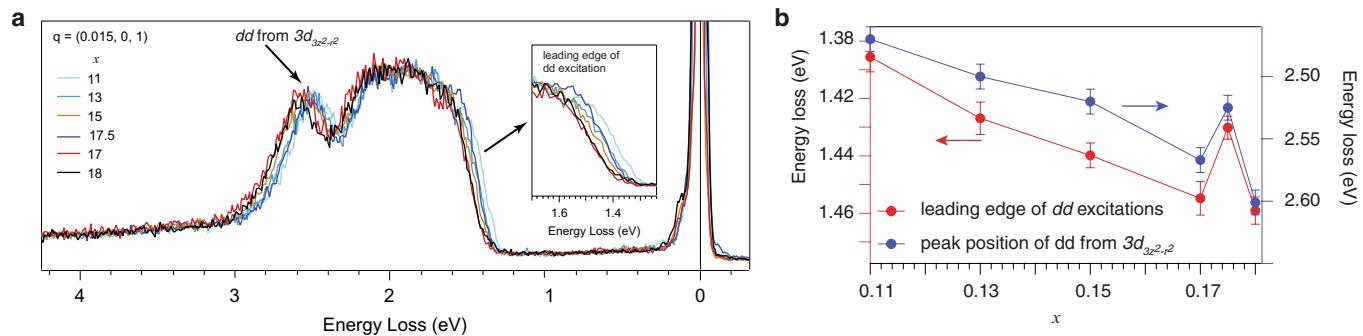
Extended Data Fig. 3 | Three-dimensionality of the zone centre excitations in NCCO. **a, b**, RIXS intensity maps of NCCO ($x = 0.15$) for momentum transfer along the h -direction at $l^* = 0.5$ and $l^* = 0.825$. Red and grey symbols indicate least-squares-fit peak positions of the zone centre excitation and the paramagnon, respectively. The inset indicates the probe direction in reciprocal space. **c**, RIXS intensity map of NCCO

($x = 0.15$) for momentum transfer along the out-of-plane direction at $h = 0.025$. White symbols indicate fitted peak positions of the zone centre excitation. Error bars are estimated from the uncertainty in energy-loss reference-point determination (± 0.01 eV) together with the standard deviation of the fits.



Extended Data Fig. 4 | Fits of the plasmon dispersion in the layered electron gas model. **a.** Fits (solid lines) of the mode energies of LCCO ($x = 0.175$) (red symbols) as a function of in-plane momentum transfer q_{\parallel} along the h -direction at $l^* = 0.5$, $l^* = 0.825$ and $l^* = 0.9$. The fit is global,

that is, the three l^* datasets are fitted simultaneously with the same fit parameter, as described in the Methods. Error bars of the data points are the same as those estimated in Fig. 2c.



Extended Data Fig. 5 | Verification of electron doping systematics via dd excitations in the RIXS spectra. a, dd excitations in RIXS spectra at momentum transfer $(0.015, 0, 1)$ taken from samples with different Ce doping concentrations x . The energy positions of dd excitations shift to higher energy with increasing electron doping, which can be used as an internal reference to verify the doping concentrations. The inset shows a zoom-in of the leading-edge region of the dd excitations. **b,** The correlation

between the Ce concentration x and the energy of the dd-leading edge (inflection point) and the $3d_{z^2-r^2}$ peak. All samples show good correlation except for the $x = 0.175$ sample, indicating a larger uncertainty of its doping concentration. Thus, the $x = 0.175$ data were not included in Fig. 4. The error bars are estimated from the standard deviation of the fit used to determine the energy of the dd-leading edge and the $3d_{z^2-r^2}$ peak.

Catalytic enantioconvergent coupling of secondary and tertiary electrophiles with olefins

Zhaobin Wang¹, Haolin Yin¹ & Gregory C. Fu^{1,*}

Carbon–carbon bonds, including those between sp^3 -hybridized carbon atoms (alkyl–alkyl bonds), typically comprise much of the framework of organic molecules. In the case of sp^3 -hybridized carbon, the carbon can be stereogenic and the particular stereochemistry can have implications for structure and function^{1–3}. As a consequence, the development of methods that simultaneously construct alkyl–alkyl bonds and control stereochemistry is important, although challenging. Here we describe a strategy for enantioselective alkyl–alkyl bond formation, in which a racemic alkyl electrophile is coupled with an olefin in the presence of a hydrosilane, rather than via a traditional electrophile–nucleophile cross-coupling, through the action of a chiral nickel catalyst. We demonstrate that families of racemic alkyl halides—including secondary and tertiary electrophiles, which have not previously been shown to be suitable for enantioconvergent coupling with alkyl metal nucleophiles—cross-couple with olefins with good enantioselectivity and yield under very mild reaction conditions. Given the ready availability of olefins, our approach opens the door to developing more general methods for enantioconvergent alkyl–alkyl coupling.

The transition-metal-catalysed enantioconvergent cross-coupling of a readily available racemic secondary alkyl electrophile with an alkyl metal nucleophile is an effective strategy for addressing the twofold task of alkyl–alkyl bond formation and controlling enantioselectivity (Fig. 1a); however, so far, methods that proceed with high enantioselectivity and good yield have been described for only a small fraction of the possible permutations of electrophiles and nucleophiles^{4–7}. Consequently, the development of alternative strategies for achieving enantioconvergent substitution reactions of racemic electrophiles could have a substantial effect on organic synthesis. Here, we demonstrate that the reductive coupling of a racemic alkyl halide with an olefin^{8,9} complements previous approaches (Fig. 1b). Some noteworthy advantages of our strategy are: olefins are typically more attractive coupling partners than are alkyl metal reagents; the single catalyst described here is more versatile than any single catalyst yet described for electrophile–nucleophile coupling, enabling highly enantioselective cross-coupling of racemic tertiary electrophiles and a wide variety of secondary electrophiles; and the coupling conditions are mild.

Carbonyl groups that bear an α stereocentre occur in various bioactive compounds^{10,11}. As a result, the development of methods to generate such stereocentres in highly enantioenriched form is important. Whereas early efforts concentrated largely on the use of stoichiometric chiral auxiliaries to control the desired stereochemistry, recent studies have increasingly focused on asymmetric catalysis^{12,13}, including enantioconvergent alkyl–alkyl cross-coupling (electrophile–nucleophile)¹⁴.

Racemic secondary α -haloamides, which bear an acidic proton, have not been reported to be suitable partners in enantioselective electrophile–nucleophile cross-coupling. We determined that a chiral nickel-bis(oxazoline) catalyst achieves the enantioconvergent coupling of various amides and olefins, providing the desired products with generally high enantiomeric excess and good yield (Fig. 2a, 1–19). The observation of high enantioselectivity and good yield (for example, 1: 94% enantiomeric excess, 84% yield) when the racemic electrophile is

the limiting reagent establishes that both enantiomers of the electrophile are being converted into the enantioenriched product; that is, this is an enantioconvergent reaction, not a simple kinetic resolution. From a practical point of view it is noteworthy that the coupling proceeds in only slightly diminished yield (and with no loss in enantiomeric excess) when run under an atmosphere of air, in the presence of one equivalent of water, with less catalyst or with less olefin (66%–78% yield; Fig. 2a, 1). Furthermore, the coupling can be carried out on a gram scale with similar efficiency (94% enantiomeric excess, 88% yield).

The scope of this enantioconvergent cross-coupling is fairly broad with respect to the R substituent on the α -carbon of the electrophile (Fig. 2a, 1–4) and on the olefin (5–14), providing the desired alkyl–alkyl coupling product in good enantiomeric excess and yield. A diverse array of functional groups, including an unactivated primary alkyl chloride and bromide, an acetal, an ester, a carbamate, a boronate ester and an ether, are compatible with the reaction conditions (Fig. 2a), as is an alcohol, an aldehyde, an aryl iodide, a benzofuran, an epoxide, an indole, a ketone, a secondary amine and a thioether (Supplementary Information). The stereochemistry of the chiral catalyst, rather than existing stereocentres on the olefin coupling partner, predominantly determines the stereochemistry of the coupling products 11–14.

The scope with respect to the electrophile is not limited to changes in the α substituent of the original secondary amide (Fig. 2a, 1–4); the carbonyl group can also be altered (15–26). Thus, the same chiral nickel-bis(oxazoline) catalyst effects the cross-coupling of an array of racemic secondary

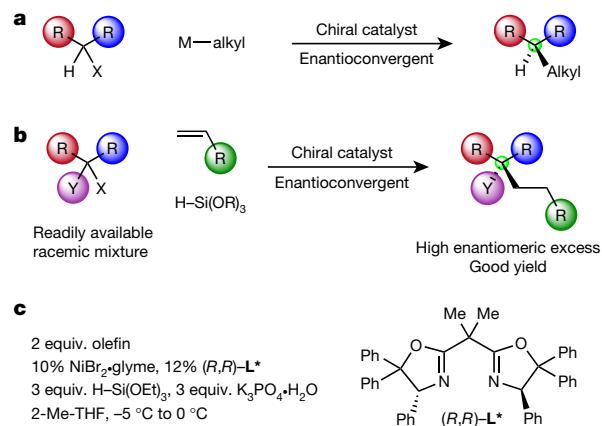
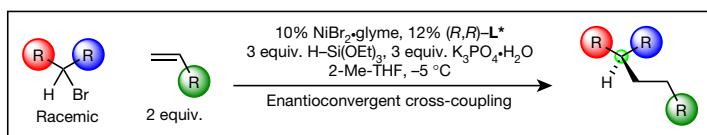
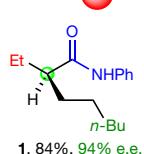
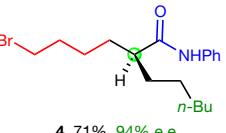
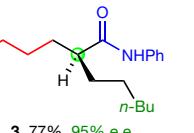
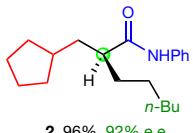
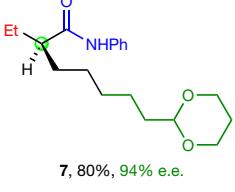
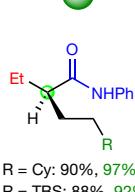


Fig. 1 | Transition-metal-catalysed enantioconvergent alkyl–alkyl cross-coupling reactions of racemic alkyl electrophiles. **a**, Previous approach: electrophile–nucleophile cross-coupling, which was limited to cross-coupling between secondary electrophiles and alkyl metals. **b**, Our approach: electrophile–olefin cross-coupling, which enables cross-coupling between secondary and tertiary electrophiles and olefins. The advantages of this approach are: olefins are widely available as coupling partners; the method is versatile, being effective for diverse (secondary and tertiary) electrophiles; and we use mild reaction conditions, with good functional-group compatibility. **c**, General method for our reaction. R, carbon substituent; X, leaving group; M, metal; Y, R or H.

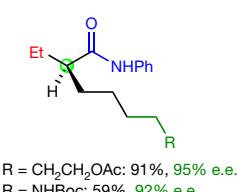
¹Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, USA. *e-mail: gcfu@caltech.edu

**a**Variation of R^1 

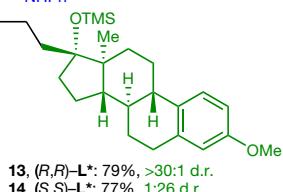
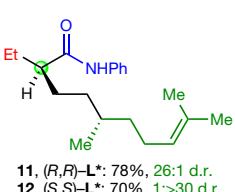
Under air: 66%, 94% e.e.
1 equiv. H_2O added: 78%, 95% e.e.
5% NiBr_2 -glyme, 6% (R,R)-L*: 75%, 94% e.e.
1.5 equiv. olefin: 70%, 94% e.e.
Gram-scale coupling: 88%, 94% e.e.

Variation of R^2 

5, $\text{R} = \text{Cy}$: 90%, 97% e.e.
6, $\text{R} = \text{TBS}$: 88%, 92% e.e.

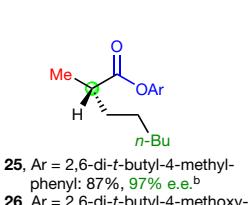
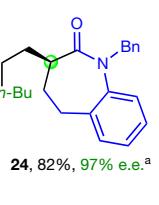
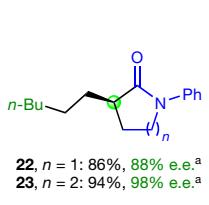
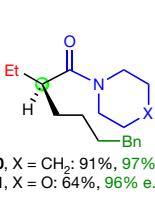
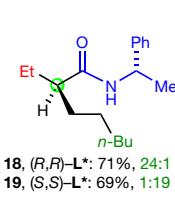
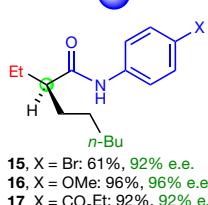


7, $\text{R} = \text{CH}_2\text{CH}_2\text{OAc}$: 91%, 95% e.e.
9, $\text{R} = \text{NHBOC}$: 59%, 92% e.e.
10, $\text{R} = \text{Bpin}$: 64%, 94% e.e.



11, (R,R)-L*: 78%, 26:1 d.r.
12, (S,S)-L*: 70%, 1:>30 d.r.

13, (R,R)-L*: 79%, >30:1 d.r.
14, (S,S)-L*: 77%, 1:26 d.r.

Variation of R^3 

15, $\text{X} = \text{Br}$: 61%, 92% e.e.
16, $\text{X} = \text{OMe}$: 96%, 96% e.e.
17, $\text{X} = \text{CO}_2\text{Et}$: 92%, 92% e.e.

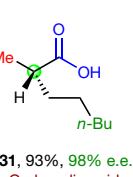
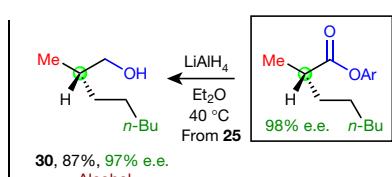
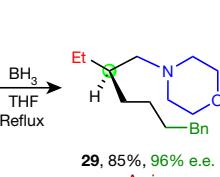
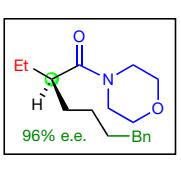
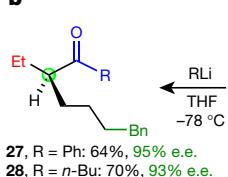
18, (R,R)-L*: 71%, 24:1 d.r.
19, (S,S)-L*: 69%, 1:19 d.r.

20, $\text{X} = \text{CH}_2$: 91%, 97% e.e.
21, $\text{X} = \text{O}$: 64%, 96% e.e.

22, $n = 1$: 86%, 88% e.e.^a
23, $n = 2$: 94%, 98% e.e.^a

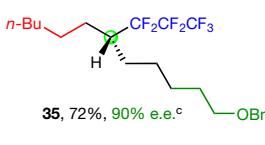
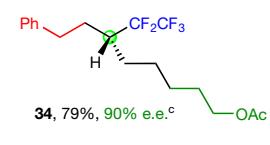
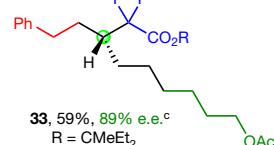
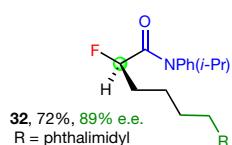
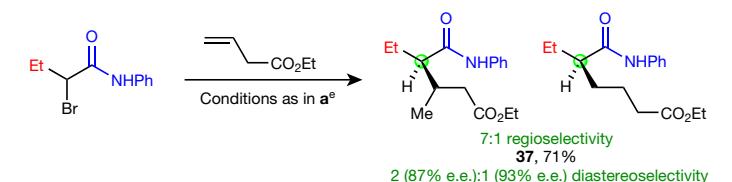
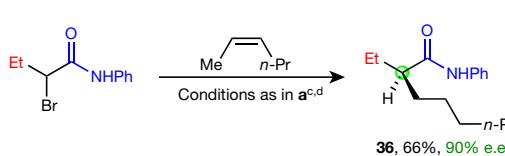
24, 82%, 97% e.e.^a
25, $\text{Ar} = 2,6\text{-di-}t\text{-butyl-4-methylphenyl}$: 87%, 97% e.e.^b

26, $\text{Ar} = 2,6\text{-di-}t\text{-butyl-4-methoxyphenyl}$: 86%, 98% e.e.^b

b

27, $\text{R} = \text{Ph}$: 64%, 95% e.e.
28, $\text{R} = \text{n-Bu}$: 70%, 93% e.e.

Aryl and alkyl ketones

c**d**

37, 71%
2 (87% e.e.): 1 (93% e.e.) diastereoselectivity

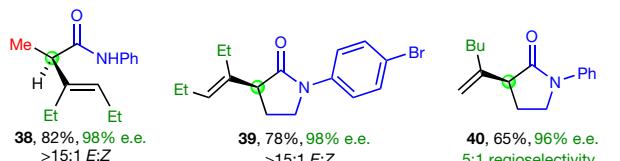
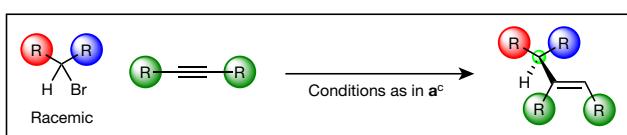
e

Fig. 2 | Enantioconvergent alkyl–alkyl cross-coupling of racemic secondary alkyl electrophiles with olefins. **a**, Secondary α -halocarbonyl compounds as electrophiles, including the effect of reaction parameters on enantioselectivity and yield (yellow shaded box). **b**, Transformation into other families of enantioenriched compounds. **c**, Other families of electrophiles. **d**, Chain-walking and directed alkylation. **e**, Alkynes as

coupling partners. ^aIodide as the leaving group. ^bReaction run in toluene at room temperature. ^cReaction run in toluene. ^dReaction run with 2 equiv. NaI . ^eReaction run with 0.5 equiv. $(n\text{-Bu})_4\text{NI}$. RT, room temperature; d.r., diastereomeric ratio; e.e., enantiomeric excess. All data are the average of two experiments.

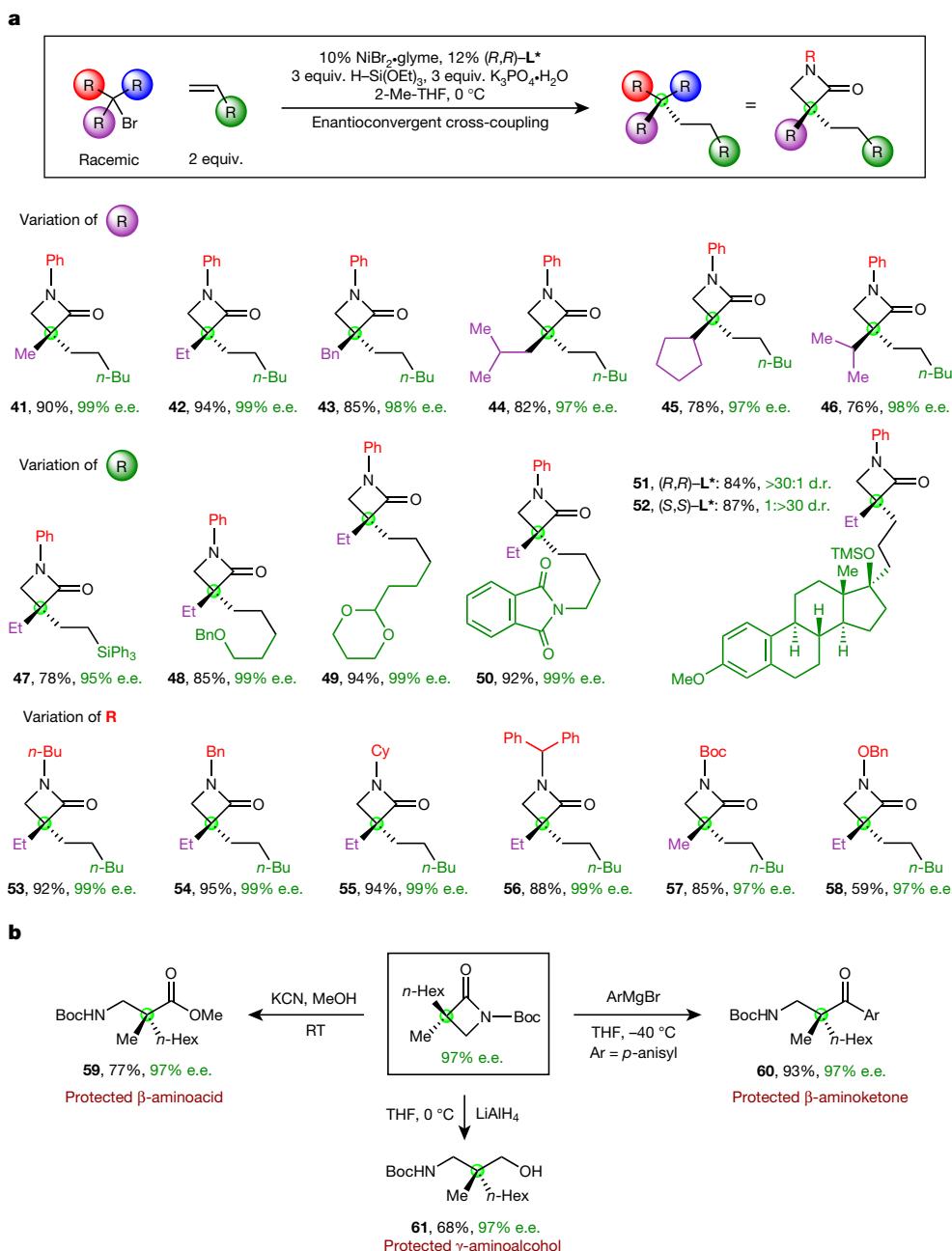


Fig. 3 | Enantioconvergent alkyl–alkyl cross-coupling of racemic tertiary alkyl electrophiles with olefins. a, Couplings. b, Transformation into other families of enantioenriched compounds. All data are the average of two experiments.

amides (15–19), tertiary amides (20 and 21), lactams (22–24) and esters (25 and 26) with generally good enantiomeric excess and yield. Other than tertiary amides¹⁴, none of these families of electrophiles has previously been shown to participate in enantioconvergent alkyl–alkyl coupling with an alkyl metal reagent as the nucleophile. In the case of a chiral amide, the stereochemistry of the catalyst primarily dictates the stereochemistry α to the carbonyl group in the cross-coupling product (18 and 19). The products of these enantioconvergent alkyl–alkyl coupling reactions can be converted in a single step without substantial racemization to other important families of enantioenriched compounds, including aryl and alkyl ketones, amines, alcohols and carboxylic acids (Fig. 2b, 27–31).

Furthermore, the chiral nickel–bis(oxazoline) catalyst can achieve enantioconvergent cross-coupling of very different families of racemic alkyl electrophiles to generate highly enantioenriched target structures. There is growing interest in fluorinated compounds in medicinal chemistry¹⁵. When the alkyl substituent of an α -halocarbonyl electrophile is replaced with a fluoro substituent, the catalyst effects selective

substitution of the bromide in the presence of the fluoride, providing the desired product with good enantioselectivity (Fig. 2c, 32: 89% enantiomeric excess). Furthermore, the catalyst achieves asymmetric alkyl–alkyl bond formation not only when the bromide leaving group is α to a carbonyl group, but also when it is β (33). Finally, the carbonyl group can be removed entirely (34 and 35). These results demonstrate the potential of our approach to enantioconvergent alkyl–alkyl coupling of racemic alkyl electrophiles: when alkyl metal reagents have previously been used as the cross-coupling partner, no single chiral catalyst has been shown to be effective for such a diverse range of secondary alkyl electrophiles^{4,5}.

In the case of a 1,2-disubstituted olefin, the nickel catalyst is able to chain-walk to achieve exclusive n -alkylation (Fig. 2d, 36^{16,17}; see below for a mechanistic discussion). On the other hand, the presence of a suitably positioned directing group^{18,19} can reverse the general preference of the catalyst for n -alkylation, leading primarily to the branched product (37).

One of the most important challenges in the field of enantioconvergent alkyl–alkyl cross-coupling is the development of

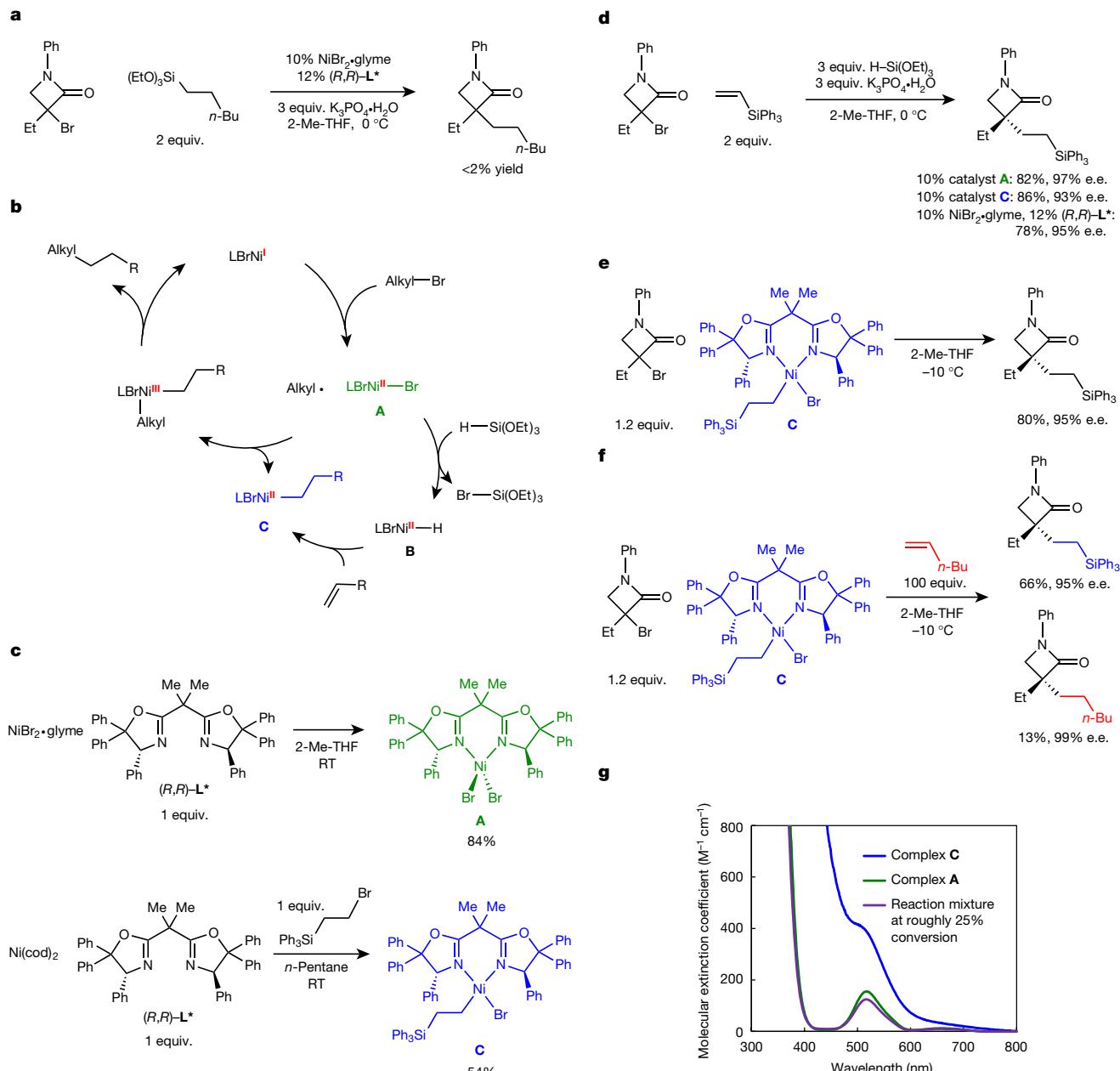


Fig. 4 | Mechanism. **a**, Evidence against a conventional cross-coupling mechanism: the proposed intermediate is not chemically competent. **b**, Possible mechanism for enantioconvergent electrophile–olefin cross-coupling. For the sake of simplicity, all steps are drawn as irreversible; H-Si(OEt)_3 represents the hydrosilane activated by $\text{K}_3\text{PO}_4 \cdot \text{H}_2\text{O}$. **c**, Synthesis of proposed intermediates **A** (top) and **C** (bottom). **d**, Competence of

methods to couple racemic tertiary electrophiles with high enantioselectivity and yield to generate quaternary stereocentres²⁰. Success so far has largely been restricted to the use of enolate nucleophiles or allylic coupling partners^{21–23}. By contrast, we determined that the same nickel-bis(oxazoline) catalyst that is effective for enantioconvergent coupling of racemic secondary alkyl halides (Fig. 2) can be applied to corresponding reactions of tertiary electrophiles, specifically, α -halo- β -lactams, affording the desired products in excellent enantiomeric excess and generally good yield (Fig. 3a). Such cross-coupling of a tertiary alkyl electrophile with an olefin has not previously been described, even to generate a racemic product.

As illustrated in Fig. 3a, high enantiomeric excess is observed in enantioconvergent alkyl–alkyl coupling of tertiary electrophiles that bear various substituents. For example, the α -alkyl group of the racemic α -halo- β -lactam ranges in steric demand from methyl to isopropyl

complexes **A** and **C** as catalysts. **e**, Chemical competence of complex **C**. **f**, Support for β -migratory insertion (**B** \rightarrow **C**): incorporation of 1-hexene into the cross-coupling product. **g**, Identification and quantification of the likely resting state via ultraviolet-visible spectroscopy. The purple trace represents a reaction at roughly 25% conversion, using the coupling partners illustrated in **d** with 10% $\text{NiBr}_2 \cdot \text{glyme}$ and 12% $(R,R)-\text{L}^*$ as the catalyst.

(41–46). Many terminal olefins are suitable coupling partners, including substrates that contain a silane, an ether, an acetal, an imide and a steroid subunit (47–52). Furthermore, excellent enantioselectivity is observed regardless of whether the substituent on the nitrogen of the β -lactam is an aryl, alkyl, Boc or alkoxy group (53–58).

β -Lactams, including compounds with an α quaternary centre, are important not only as endpoints^{24–27}, but also as intermediates in asymmetric synthesis²⁸. β -Aminoacids, β -aminoketones and γ -aminoalcohols—all of which are important targets owing to their occurrence as subunits in bioactive compounds such as dexmethylphenidate, tolperisone and propranolol—can be generated from β -lactams without loss of enantiomeric excess (Fig. 3b, 59–61).

We have begun to investigate the mechanism of the catalytic enantioconvergent alkyl–alkyl cross-coupling process. One possible pathway is that the olefin undergoes nickel-catalysed hydrosilylation²⁹ and that the resulting

alkylsilane serves as a nucleophile in a conventional electrophile–nucleophile (Hiyama-type) cross-coupling. To test this possibility, we subjected the putative alkylsilane to the reaction conditions; however, we observed essentially no cross-coupling, which rules out this pathway (Fig. 4a).

Our current hypothesis is that enantioconvergent electrophile–olefin cross-coupling proceeds through the pathway illustrated in Fig. 4b, which builds on a mechanistic study of an enantioconvergent electrophile–nucleophile cross-coupling³⁰. $\text{LBrNi}^{\text{II}}\text{-Br}$ (**A**) reacts with the hydrosilane to generate a nickel–hydride complex ($\text{LBrNi}^{\text{II}}\text{-H}$; **B**). Olefin complexation followed by β -migratory insertion then results in a nickel–alkyl complex ($\text{LBrNi}^{\text{II}}\text{-CH}_2\text{CH}_2\text{R}$; **C**), which enters the reaction cycle for electrophile–nucleophile cross-coupling.

Using electron paramagnetic resonance spectroscopy, we examined a cross-coupling in progress. We observed no signal, which is consistent with the resting state of the catalytic cycle being one or more $\text{Ni}(\text{II})$ complexes, such as **A–C**, rather than $\text{Ni}(\text{I})$ and $\text{Ni}(\text{III})$ complexes (Fig. 4b). We independently synthesized $\text{Ni}(\text{II})$ complexes **A** and **C** (Fig. 4c) and crystallographically characterized complex **A**.

Our studies of complexes **A** and **C** are consistent with the mechanism illustrated in Fig. 4b. When used in place of $\text{NiBr}_2\text{-glyme}/(R,R)\text{-L}^*$, both complexes furnish the cross-coupling product with enantiomeric excess and yield similar to those achieved when using the catalyst generated in situ under our standard conditions (Fig. 4d). Furthermore, nickel–alkyl complex **C** is chemically competent, reacting with an electrophile in a stoichiometric coupling to provide the product with the expected enantioselectivity and in good yield (Fig. 4e; this process is inhibited by the addition of TEMPO, a radical trap³¹); we anticipate that a $\text{Ni}(\text{I})$ complex (generated by reductive elimination of $\text{Ni}(\text{II}) \rightarrow \text{Ni}(0)$ followed by com-proportionation with $\text{Ni}(\text{II})$ to form $\text{Ni}(\text{I})$) initiates this reaction³⁰. When this reaction is conducted in the presence of an olefin (1-hexene), we observe a mixture of two alkylation products (Fig. 4f). This finding highlights the accessibility of nickel hydride **B** and its β -migratory insertion, which are key aspects of our proposed mechanism (Fig. 4b). Finally, analysis via ultraviolet–visible spectroscopy of a coupling reaction in progress is consistent with the suggestion that about 80% of the nickel in the reaction mixture is present as complex **A** (Fig. 4g; no evidence for complex **C**).

In summary, we have described a strategy for enantioconvergent alkyl–alkyl cross-coupling, specifically, nickel-catalysed coupling of racemic secondary and tertiary alkyl electrophiles with olefins. Asymmetric electrophile–olefin cross-coupling has considerable advantages over traditional electrophile–nucleophile coupling, owing to the ready availability of olefins, the broad scope and mild reaction conditions. We anticipate that our strategy will be applicable to enantioconvergent cross-coupling with various unsaturated compounds other than olefins (see, for example, Fig. 2e), thereby opening the door to new methods for asymmetric catalysis.

Data availability

The data that support the findings of this study are available within the paper, its Supplementary Information (experimental procedures and characterization data) and from the Cambridge Crystallographic Data Centre (<https://www.ccdc.cam.ac.uk/structures>; crystallographic data are available free of charge under CCDC reference numbers 1822790–1822793, 1839344–1839346 and 1861568).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0669-y>.

Received: 30 April 2018; Accepted: 11 October 2018;

Published online 18 October 2018.

- Carreira, E. M. & Yamamoto, H. (eds) *Comprehensive Chirality* (Academic, Amsterdam, 2012).
- Lin, G.-Q., You, Q.-D. & Cheng, J.-F. (eds) *Chiral Drugs: Chemistry and Biological Action* (Wiley, New York, 2011).
- Lovering, F., Bikker, J. & Humblet, C. Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **52**, 6752–6756 (2009).
- Choi, J. & Fu, G. C. Transition metal-catalyzed alkyl–alkyl bond formation: another dimension in cross-coupling chemistry. *Science* **356**, eaaf7230 (2017).

- Fu, G. C. Transition-metal catalysis of nucleophilic substitution reactions: a radical alternative to $\text{S}_{\text{N}}1$ and $\text{S}_{\text{N}}2$ processes. *ACS Cent. Sci.* **3**, 692–700 (2017).
- Iwasaki, T. & Kambe, N. in *Comprehensive Organic Synthesis* 2nd edn, Vol. 3 (eds Knochel, P. et al.) 337–391 (Elsevier, Amsterdam, 2014).
- Geist, E., Kirschning, A. & Schmidt, T. $\text{sp}^3\text{-sp}^3$ coupling reactions in the synthesis of natural products and biologically active molecules. *Nat. Prod. Rep.* **31**, 441–448 (2014).
- Lu, X. et al. Practical carbon–carbon bond formation from olefins through nickel-catalysed reductive olefin hydrocarbonation. *Nat. Commun.* **7**, 11129 (2016).
- Wang, Y.-M., Bruno, N. C., Placeres, A. L., Zhu, S. & Buchwald, S. L. Enantioselective synthesis of carbo- and heterocycles through a CuH-catalyzed hydroalkylation approach. *J. Am. Chem. Soc.* **137**, 10524–10527 (2015).
- Tobert, J. A. Lovastatin and beyond: the history of the HMG-CoA reductase inhibitors. *Nat. Rev. Drug Discov.* **2**, 517–526 (2003).
- Ganellin, C. R. in *Introduction to Biological and Small Molecule Drug Research and Development* (eds Ganellin, C. R. et al.) 339–416 (Elsevier, Amsterdam, 2013).
- Stoltz, B. M. et al. in *Comprehensive Organic Synthesis* 2nd edn, Vol. 3 (eds Knochel, P. et al.) 1–55 (Elsevier, Amsterdam, 2014).
- MacMillan, D. W. C. & Watson, A. J. B. in *Science of Synthesis: Stereoselective Synthesis* Vol. 3 (ed. Evans, P. A.) 675–745 (Thieme, New York, 2011).
- Fischer, C. & Fu, G. C. Asymmetric nickel-catalyzed Negishi cross-couplings of secondary α -bromo amides with organozinc reagents. *J. Am. Chem. Soc.* **127**, 4594–4595 (2005).
- Gouverneur, V. & Müller, K. *Fluorine in Pharmaceutical and Medicinal Chemistry* (Imperial College Press, London, 2012).
- Juliá-Hernández, F., Moragas, T., Cornellà, J. & Martin, R. Remote carboxylation of halogenated aliphatic hydrocarbons with carbon dioxide. *Nature* **545**, 84–88 (2017).
- Zhou, F., Zhu, J., Zhang, Y. & Zhu, S. NiH-catalyzed reductive relay hydroalkylation: a strategy for the remote $\text{C}(\text{sp}^3)\text{-H}$ alkylation of alkenes. *Angew. Chem. Int. Ed.* **57**, 4058–4062 (2018).
- Hoveyda, A. H., Evans, D. A. & Fu, G. C. Substrate-directable chemical reactions. *Chem. Rev.* **93**, 1307–1370 (1993).
- Derosa, J., Tran, V. T., Boulous, M. N., Chen, J. S. & Engle, K. M. Nickel-catalyzed β,γ -dicarbofunctionalization of alkyl carbonyl compounds via conjunctive cross-coupling. *J. Am. Chem. Soc.* **139**, 10657–10660 (2017).
- Quasdorff, K. W. & Overman, L. E. Catalytic enantioselective synthesis of quaternary carbon stereocentres. *Nature* **516**, 181–191 (2014).
- Ding, C.-H. & Hou, X.-L. in *Comprehensive Organic Synthesis* 2nd edn, Vol. 4 (eds Knochel, P. et al.) 648–698 (Elsevier, Amsterdam, 2014).
- Murakata, M., Jono, T., Mizuno, Y. & Hoshino, O. Construction of chiral quaternary carbon centers by catalytic enantioselective radical-mediated allylation of α -iodolactones using allyltributyltin in the presence of a chiral Lewis acid. *J. Am. Chem. Soc.* **119**, 11713–11714 (1997).
- Ma, S., Han, X., Krishnan, S., Virgil, S. C. & Stoltz, B. M. Catalytic enantioselective stereoablative alkylation of 3-haloindoles: facile access to oxindoles with C3 all-carbon quaternary stereocenters. *Angew. Chem. Int. Ed.* **48**, 8037–8041 (2009).
- Banik, B. K. (ed.) *β -Lactams: Unique Structures of Distinction for Novel Molecules* (Springer, Berlin, 2013).
- Decuyper, L. et al. Antibacterial and β -lactamase inhibitory activity of monocyclic β -lactams. *Med. Res. Rev.* **38**, 426–503 (2018).
- Galletti, P. & Giacomini, D. Monocyclic β -lactams: new structures for new biological activities. *Curr. Med. Chem.* **18**, 4265–4283 (2011).
- Chrusciel, R. A. et al. Therapeutic compounds and compositions. WO patent 2015/120062 A2 (2015).
- Ojima, I., Zuniga, E. S. & Seitz, J. D. in *β -Lactams: Unique Structures of Distinction for Novel Molecules* (ed. Banik, B. K.) 1–64 (Springer, Berlin, 2013).
- Du, X. & Huang, Z. Advances in base-metal-catalyzed alkene hydroisilylation. *ACS Catal.* **7**, 1227–1243 (2017).
- Schley, N. D. & Fu, G. C. Nickel-catalyzed Negishi arylations of propargylic bromides: a mechanistic investigation. *J. Am. Chem. Soc.* **136**, 16588–16593 (2014).
- Henry-Riyad, H. et al. in *Handbook of Reagents for Organic Synthesis* (ed. Fuchs, P. L.) 620–626 (Wiley, New York, 2013).

Acknowledgements Support has been provided by the National Institutes of Health (National Institute of General Medical Sciences, R01-GM262871) and the Gordon and Betty Moore Foundation (Caltech Center for Catalysis and Chemical Synthesis). We thank L. M. Henling, D. G. VanderVelde and S. C. Virgil for assistance and discussions.

Reviewer information *Nature* thanks C. Gosmini and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Z.W. and H.Y. performed all experiments. Z.W. and G.C.F. wrote the manuscript. All authors contributed to the analysis and the interpretation of the results.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0669-y>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to G.C.F.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Urbanization exacerbated the rainfall and flooding caused by hurricane Harvey in Houston

Wei Zhang¹, Gabriele Villarini^{1*}, Gabriel A. Vecchi^{2,3} & James A. Smith⁴

Category 4 landfalling hurricane Harvey poured more than a metre of rainfall across the heavily populated Houston area, leading to unprecedented flooding and damage. Although studies have focused on the contribution of anthropogenic climate change to this extreme rainfall event^{1–3}, limited attention has been paid to the potential effects of urbanization on the hydrometeorology associated with hurricane Harvey. Here we find that urbanization exacerbated not only the flood response but also the storm total rainfall. Using the Weather Research and Forecast model—a numerical model for simulating weather and climate at regional scales—and statistical models, we quantify the contribution of urbanization to rainfall and flooding. Overall, we find that the probability of such extreme flood events across the studied basins increased on average by about 21 times in the period 25–30 August 2017 because of urbanization. The effect of urbanization on storm-induced extreme precipitation and flooding should be more explicitly included in global climate models, and this study highlights its importance when assessing the future risk of such extreme events in highly urbanized coastal areas.

North Atlantic tropical cyclones are among the costliest natural hazards both in terms of fatalities and economic impacts, with the devastation left by 2017 hurricanes Harvey, Irma and Maria typical of the havoc tropical cyclones can cause. There are multiple hazards associated with these events, including storm surge, strong winds, heavy rainfall and flooding. An analysis of 28 tropical cyclones over the 2001–2014 period found that around two-thirds of the residential flood insurance claims were caused by riverine flooding⁴, highlighting the major impact of these events for both coastal and inland communities.

The devastation caused by hurricane Harvey in Houston is a reminder of the rainfall and flooding that can be associated with these storms. Between 25 and 30 August 2017, hurricane Harvey dropped more than 1,300 mm of rain over and around Houston, leading to unprecedented flooding in large areas of the city^{1–3}. In the aftermath of this storm, different studies estimated the return period of the rainfall associated with this event and quantified the human-induced climate change signal using a combination of observations and climate models. In ref. ¹ it was found that the return period of Harvey's rainfall was around 2,000 years in the late twentieth century, and predicted to drop to 100 years by the end of this century. In ref. ² and ref. ³ it was found that human-induced climate change made this event between 1.5 and 5 times, or at least 3.5 times, more likely, respectively.

Thus, the literature on the anthropogenic contribution to hurricane Harvey has focused on precipitation and on the part that human-induced climate change may have played. Here, we seek to answer a complementary question related to the anthropogenic contribution to Harvey's flooding: to what extent did urbanization have a role in the heavy rainfall and flooding associated with hurricane Harvey? From a hydrologic perspective, increases in urbanization are expected to lead to faster runoff and larger peaks owing to the large reductions in infiltration (the amount of water the ground can absorb)^{5,6}. Houston has had the largest urban growth and the fifth-largest population growth in the United States over the period 2001–2011⁷. The increase in asphalt and concrete has led to

an increasing runoff ratio (that is, the ratio between runoff and precipitation) across many watersheds in the area, pointing to reduced infiltration and larger runoff for a given rainfall value^{8–12}. This increase in population and urbanization, combined with the flat clay terrain that characterizes this area, represents a very problematic mix from a flood perspective, despite the flood mitigation measures that have been put in place.

In addition to having a substantial impact on the hydrologic response, urbanization has the potential to directly influence the magnitude of extreme precipitation. This is a topic that has received substantial attention, particularly regarding the influence of urbanization on mesoscale convective systems, as determined through field campaigns, analysis

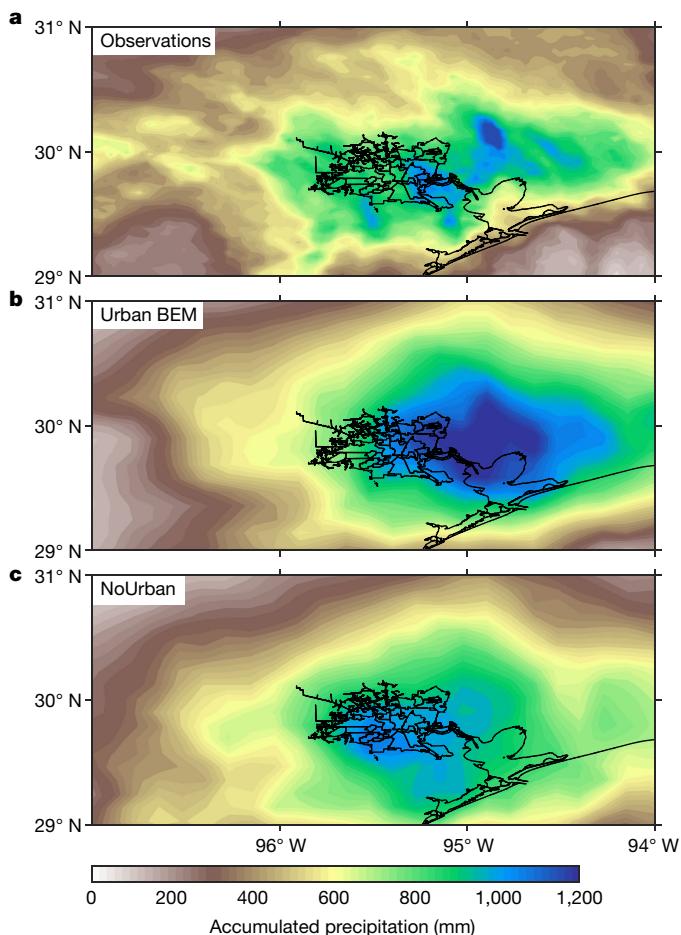


Fig. 1 | Storm total rainfall by hurricane Harvey. a–c, Accumulated precipitation for 25–30 August 2017 in observations (a), and in the 'Urban BEM' (b) and 'NoUrban' (urban land-use types replaced by croplands; c) WRF experiments. The model results represent the average of the seven members.

¹IIHR-Hydroscience & Engineering, The University of Iowa, Iowa City, IA, USA. ²Department of Geosciences, Princeton University, Princeton, NJ, USA. ³Princeton Environmental Institute, Princeton University, Princeton, NJ, USA. ⁴Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, USA. *e-mail: gabriele-villarini@uiowa.edu

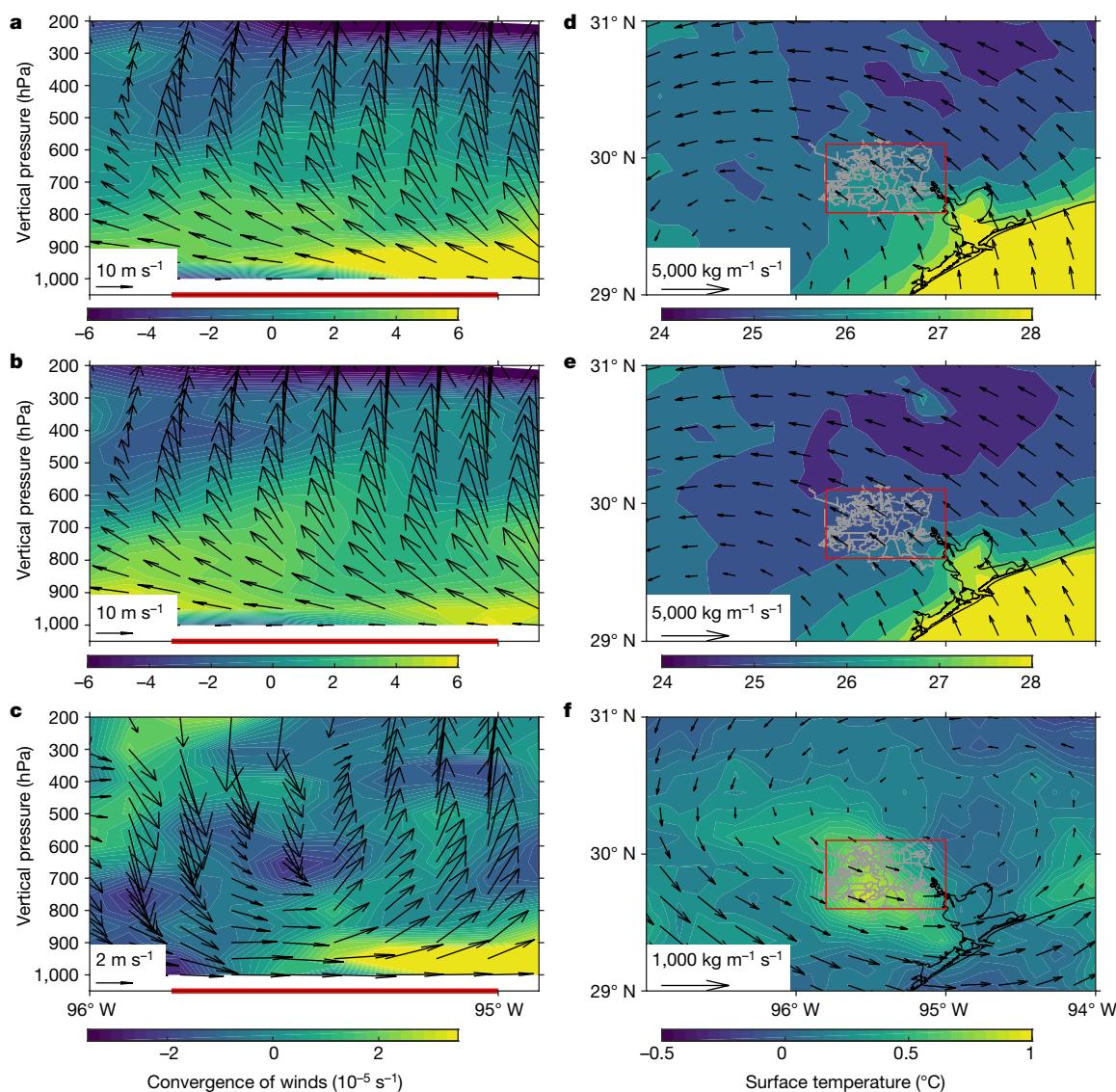


Fig. 2 | Key variables for diagnosing the effects of urbanization on hurricane Harvey. a–c, The vertical profile of wind convergence (colour scale) and zonal and vertical wind speed and direction (vertical wind speed shown here has been multiplied by 100, indicated by arrows) averaged over Houston (red rectangles in d–f) for 25–30 August 2017 for the ‘Urban BEM’ (a) and ‘NoUrban’ (b) experiments and their differences

of observations and numerical modelling. Much of the research on the topic has, however, focused on convective systems and how the urban area affects the trajectory of the convective cells^{13–15}. Much less is known regarding the urban effects on the organized tropical rainfall of a hurricane, in particular during one like hurricane Harvey, which stalled for several days. Overall, the role of the city in altering both the rainfall and runoff responses has received very limited attention, especially in the context of tropical cyclones.

Because of the substantial hydrometeorological impacts of urbanization, here we want to quantify the part that urban development of Houston has played during hurricane Harvey by accounting for both changes in total rainfall and flood response. We start by examining the role of urbanization in terms of storm total rainfall; these analyses are then followed by a focus on the hydrologic response, and we conclude our study by quantifying the impact of Houston on the flooding from hurricane Harvey. Combined with the assessments of the climate-mediated anthropogenic impacts on hurricane Harvey^{1–3}, this work will enable a fuller assessment of the total anthropogenic contribution to the flooding from this storm.

(c). The red lines at the bottom represent the zonal extent of Houston. d–f, The daily averaged vertically integrated moisture flux (indicated by arrows) and surface temperature (colour scale) for 25–30 August 2017 for the ‘Urban BEM’ (d) and ‘NoUrban’ (e) experiments with WRF and their differences (f).

To explore urban and hydrometeorological impacts on rainfall during hurricane Harvey, we performed two sets of seven-member experiments with the WRF model (see Methods). In the first set (‘Urban BEM’), we simulate hurricane Harvey by using WRF coupled with the Noah land surface model and the multi-layer building energy model (BEM). The second set of experiments (‘NoUrban’) is performed with the same setting as the first one, but with croplands replacing urban land-use types. The differences in rainfall between the two experiments represent the impacts of urbanization on Harvey’s total rainfall. During 25–30 August 2017, hurricane Harvey was responsible for more than 1,000 mm of rainfall over the Houston area, in particular over the southeastern part of the city (Fig. 1a). Overall, the ‘Urban BEM’ experiments capture reasonably well many aspects of the observed rainfall during this period (Fig. 1b). From a quantitative perspective, the rainfall totals during Harvey in the ‘Urban BEM’ experiments are comparable with those observed (Fig. 1a, b). In contrast, the ‘NoUrban’ experiments produce much less precipitation than the ‘Urban BEM’ ones, with the rainfall maxima shifted west with respect to the ‘Urban BEM’ runs. The differences in accumulated rainfall between ‘Urban

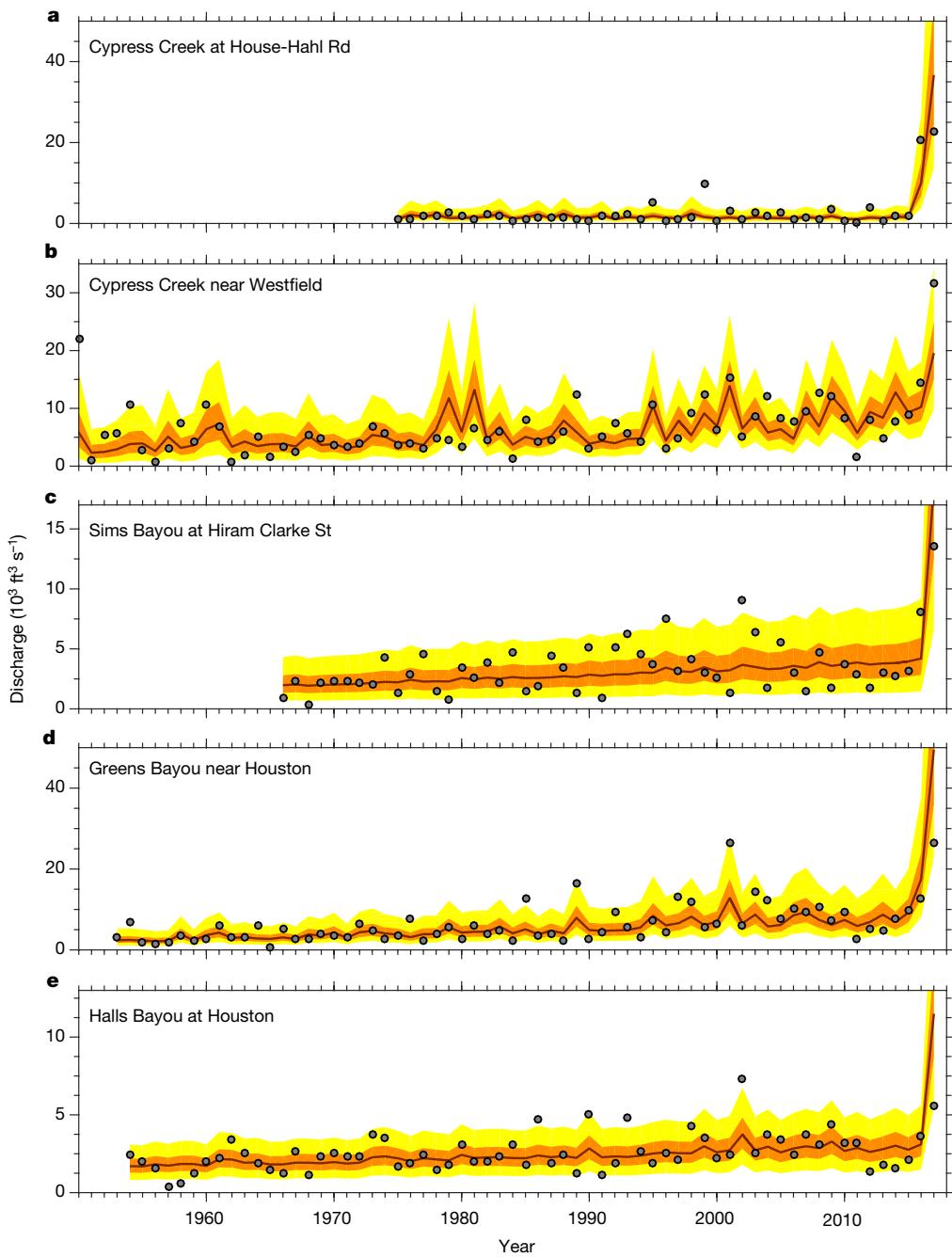


Fig. 3 | Modelling of the annual maximum peak discharge records. a–e, Modelling of the annual maximum peak discharge records at the five sites considered here (see Extended Data Fig. 5 and Extended Data

Table 1a for more information). The circles are the observations; the dark red solid lines represent the median of the fitted distribution, while the yellow (orange) areas indicate the 5th to 95th (25th to 75th) percentiles.

BEM' and 'NoUrban' experiments clearly show a large increase in rainfall arising from urbanization over the eastern part of the Houston area (Extended Data Fig. 1). Moreover, the western shift in rainfall maxima in 'NoUrban' experiments leads to more rainfall in the western part of Houston than that in the 'Urban BEM' case (Extended Data Fig. 1). These statements are also supported by statistical analyses, which indicate that these differences are statistically significant at the $P < 0.05$ level (Extended Data Fig. 1). Urbanization led to an amplification of the total rainfall along with a shift in the location of the maximum rainfall.

The enhanced rainfall in the 'Urban BEM' case and the shift of rainfall in the 'NoUrban' case are tied to the storm system's drag induced by large surface roughness. Extended Data Fig. 2 shows accumulated rainfall during 25–30 August 2017 in each member of the 'Urban BEM' and 'NoUrban' experiments. Overall, the rainfall maximum in the 'Urban BEM' experiments shifts eastward compared with the corresponding

'NoUrban' experiments for each ensemble member (Extended Data Fig. 2), similar to the ensemble mean of the results (Fig. 1) associated with the drag effects by the urban surface roughness (Extended Data Fig. 3). The consistency in the results among ensemble members lends confidence to the robustness of the impacts of urbanization in total rainfall associated with hurricane Harvey.

To further understand the physical mechanisms responsible for the enhanced rainfall due to urbanization, we analyse the vertical profile of convergence of winds and wind fields, and the 2-m surface temperature in the WRF experiments. Figure 2 shows an enhancement of the low-level convergence and of the divergence above 600 hPa in the 'Urban BEM' experiment compared to the 'NoUrban' one. The difference between these two experiments indicates enhanced low-level convergence, upper-level divergence and anomalous ascent over Houston, leading to favourable conditions for precipitation (Fig. 2c).

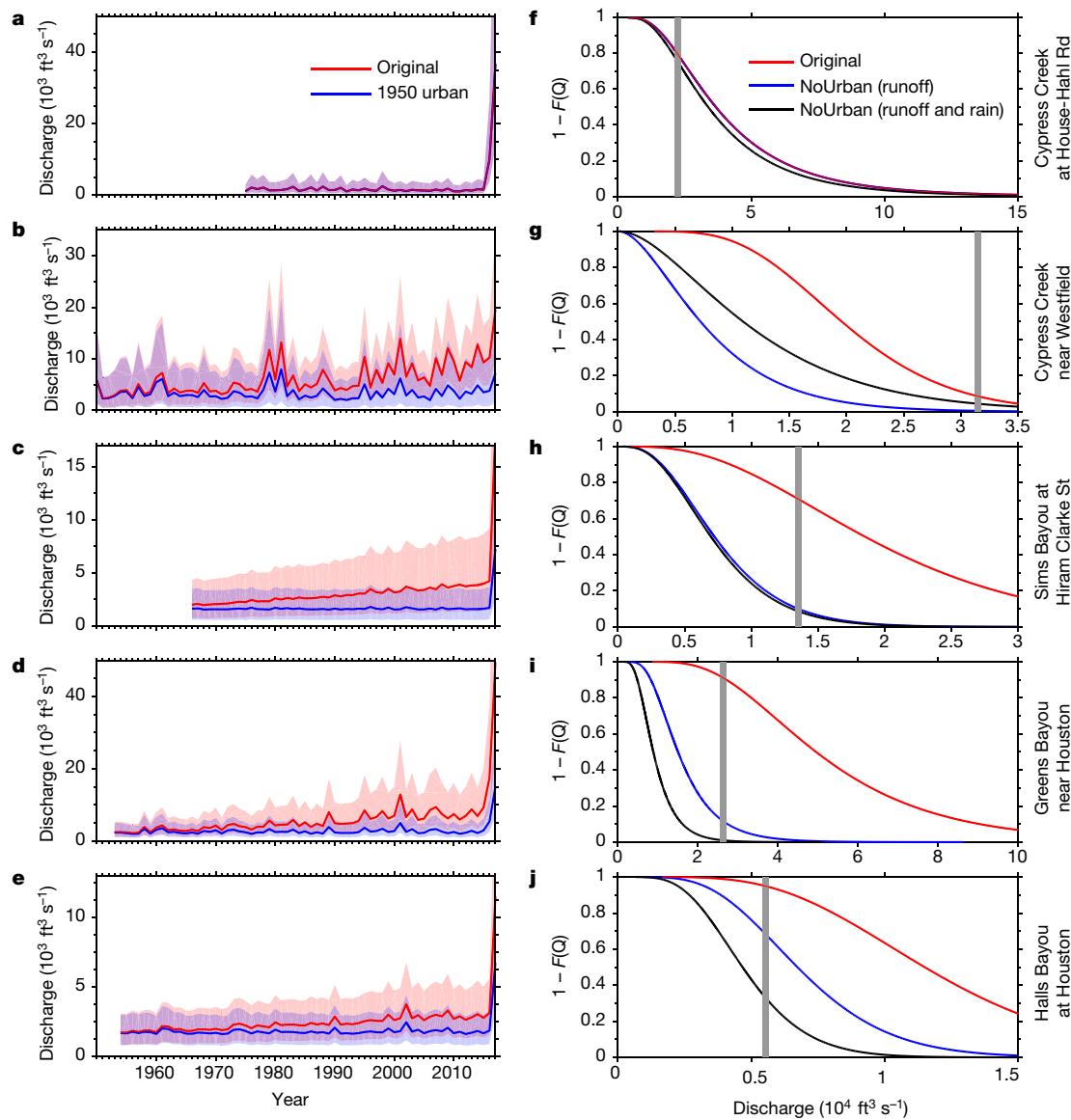


Fig. 4 | Examination of the effects of urbanization on the modelled annual maximum discharge. **a–e**, Modelled annual maximum discharge Q with original urbanization represented by population (red) and that in 1950 ('1950 urban', blue) for the five sites considered here. The shading represents the values within the 5th–95th percentile of the modelled annual maximum discharge. **f–j**, The probability ($1 - F(Q)$)

(where F is the fitted cumulative distribution function of Q) of having the value of annual maximum discharge with respect to urbanization as observed (red), with urbanization kept at the 1950 level (blue), and with urbanization kept at the 1950 level and rainfall scaled by the related change between 'Urban BEM' and 'NoUrban' simulations for each basin (black). The grey bars represent the value of annual maximum discharge for 2017.

The difference between the 'Urban BEM' and 'NoUrban' simulations highlights the presence of a cyclonic flow pattern in the moisture flux, which is also favourable for precipitation (Fig. 2d–f). Such differences in convergence and updraft are probably caused by the drag of urban surface with large roughness, associated with stronger friction velocity and roughness length in the 'Urban BEM' experiments (Extended Data Fig. 3a–f). Specifically, the friction velocity in the 'Urban BEM' simulation is markedly larger than the 'NoUrban' simulation (Extended Data Fig. 3a–c), indicating a stronger drag on the storm winds, associated with a larger surface roughness length (Extended Data Fig. 3d–f). The changes in sensible heat flux and Bowen ratio (Extended Data Fig. 3g–i) associated with urbanization and urban land-use change may lead to the destabilization of the atmosphere^{16–20}, enhancing rainfall over the eastern side of Houston. The increased urban roughness and surface warming are also associated with elevated height of the bottom boundary layer of the atmosphere, which tends to enhance precipitation (Extended Data Fig. 3j–l). These results point to the combined effects of surface drag and urban surface warming on rainfall enhancement

on the eastern side of Houston. Moreover, anthropogenic heat may influence precipitation in urban areas²¹. The 'Urban BEM' experiments can explicitly resolve anthropogenic heat to a large extent by computing anthropogenic heat using a parameterization scheme^{21,22}. Overall, rainfall over Houston in the 'Urban BEM' experiment is consistent with the results from the 'Urban BULK' experiments in which no urban canopy model is coupled with the Noah land surface model (Extended Data Fig. 4), suggesting that anthropogenic heat does not play a major part in rainfall related to hurricane Harvey.

Thus we have focused on the impacts of Houston on precipitation and provided a physical understanding of these results. However, urbanization also affects hydrologic processes, so we next quantify the role of urbanization following an approach used in ref. ²³ (see Methods). Our modelling results suggest that the year-to-year variations in annual maximum peak discharge are well captured by parsimonious statistical models relating the maximum peak discharge to precipitation and/or population (used as proxy for urbanization, see Methods) (Fig. 3). This is also supported by the goodness-of-fit diagnostics (Extended

Data Fig. 6 and Extended Data Table 2). On the basis of our modelling results, urbanization is an important predictor in controlling the magnitude and variability in the flood peak record at all the sites, with the exception of Cypress Creek at House-Hahl Road near Cypress (USGS ID 08068740), which is also the basin that can be considered as the control watershed area owing to the very limited amount of urbanization it has experienced (Extended Data Fig. 5).

To quantify the role of urbanization in terms of flooding, we use these statistical models with the observed rainfall but with urbanization kept constant at the 1950 level (that is, representative of the pre-urbanization conditions) (Fig. 4a–e). For Cypress Creek at House-Hahl Road near Cypress the results are the same, given that urbanization is not an important predictor. For the other four watersheds, there is a clear increase in the magnitude and variability associated with the urban expansion of Houston. To assess the role of hurricane Harvey, we focus on 2017 and quantify the effect of urbanization in terms of both flooding and rainfall. Figure 4f–j shows the probability of exceedance for the fitted distribution with the parameters that depend on the values of the predictors in 2017. It also shows the same results but using: (1) the value of urbanization in 1950 but the observed precipitation; and (2) the value of urbanization in 1950 and storm total rainfall scaled by the related change between ‘Urban BEM’ and ‘NoUrban’ simulations for each basin. From these results, we conclude that urbanization played an important part during this event, as shown by the shift to the left in the survival function, with changes in both hydrologic response and precipitation strongly affecting these watersheds. We also computed the risk ratio²⁴ (see Methods), which indicates that, on average, urbanization has increased the probability of an event like the flooding associated with hurricane Harvey by about 21 times (Extended Data Table 1a). Therefore, urbanization strongly exacerbated the impact that this storm has had in terms of both precipitation and flood response.

Given that hurricane winds and rainfall are projected to intensify in the future^{25,26} and urbanization is also expected to continue increasing in the major coastal cities, our work paves the way towards increasing our understanding of the risk arising from the interconnection between anthropogenic effects from urbanization and those mediated through climatic influences on hurricane-related rainfall. Planning must take into account the compounded nature of these risks, and efforts to build flood mitigation strategies must use an improved understanding of the multiple processes in place²⁷. Failure to account for urban factors would jeopardize the investment in mitigation and management of climate change impacts²⁸. Given the small scales at which an event of this kind operates in terms of urban impacts and responses, it is critical for the next generations of global climate models to be able to resolve the urban areas and their associated processes through increasing spatial resolution and improved representation of the processes that occur within the cities^{29,30}.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0676-z>.

Received: 16 April 2018; Accepted: 7 September 2018;

Published online 14 November 2018.

1. Emanuel, K. Assessing the present and future probability of hurricane Harvey's rainfall. *Proc. Natl Acad. Sci. USA* **114**, 12681–12684 (2017).
2. Jan van Oldenborgh, G. et al. Attribution of extreme rainfall from Hurricane Harvey, August 2017. *Environ. Res. Lett.* **12**, 124009 (2017).
3. Risser, M. D. & Wehner, M. F. Attributable human-induced changes in the likelihood and magnitude of the observed extreme precipitation during hurricane Harvey. *Geophys. Res. Lett.* **44**, 12457–12464 (2018).
4. Czajkowski, J., Villarini, G., Montgomery, M., Michel-Kerjan, E. & Goska, R. Assessing current and future freshwater flood risk from North Atlantic tropical cyclones via insurance claims. *Sci. Rep.* **7**, 41609 (2017).
5. Smith, J. A. et al. The regional hydrology of extreme floods in an urbanizing drainage basin. *J. Hydrometeorol.* **3**, 267–282 (2002).
6. Smith, J. A. et al. Extraordinary flood response of a small urban watershed to short-duration convective rainfall. *J. Hydrometeorol.* **6**, 599–617 (2005).
7. Bounoua, L., Nigro, J., Zhang, P., Thome, K. & Lachir, A. Mapping urbanization in the United States from 2001 to 2011. *Appl. Geogr.* **90**, 123–133 (2018).

8. Johnson, S. L. & Sayre, D. M. *Effects of Urbanization on Floods in the Houston, Texas Metropolitan Area*. Report No. 73-3, <https://pubs.er.usgs.gov/publication/wri733> (US Geological Survey, 1973).
9. Liscum, F. *Effects of Urban Development on Stormwater Runoff Characteristics for the Houston, Texas, Metropolitan Area*. Report No. 2001-4071, <https://pubs.er.usgs.gov/publication/wri014071> (US Geological Survey, 2001).
10. Khan, S. D. Urban development and flooding in Houston Texas, inferences from remote sensing data using neural network technique. *Environ. Geol.* **47**, 1120–1127 (2005).
11. Zhu, L., Quiring, S. M., Guneralp, I. & Peacock, W. G. Variations in tropical cyclone-related discharge in four watersheds near Houston, Texas. *Clim. Risk Manage.* **7**, 1–10 (2015).
12. Muñoz, L. A., Olivera, F., Giglio, M. & Berke, P. The impact of urbanization on the streamflows and the 100-year floodplain extent of the Sims Bayou in Houston, Texas. *Int. J. River Basin Manage.* **16**, 61–69 (2018).
13. Ntelekos, A. A. et al. Extreme hydrometeorological events and the urban environment: dissecting the 7 July 2004 thunderstorm over the Baltimore MD metropolitan region. *Wat. Resour. Res.* **44**, W08446 (2008).
14. Niyogi, D., Lei, M., Kishtawal, C., Schmid, P. & Shepherd, M. Urbanization impacts on the summer heavy rainfall climatology over the eastern United States. *Earth Interact.* **21**, 1–17 (2017).
15. Niyogi, D. et al. Urban modification of thunderstorms: an observational storm climatology and model case study for the Indianapolis urban region. *J. Appl. Meteorol. Climatol.* **50**, 1129–1144 (2011).
16. Oke, T. R. The energetic basis of the urban heat island. *Q. J. R. Meteorol. Soc.* **108**, 1–24 (1982).
17. Shepherd, J. M., Carter, M., Manyin, M., Messen, D. & Burian, S. The impact of urbanization on current and future coastal precipitation: a case study for Houston. *Environ. Plann. B* **37**, 284–304 (2010).
18. Baik, J.-J., Kim, Y.-H. & Chun, H.-Y. Dry and moist convection forced by an urban heat island. *J. Appl. Meteorol.* **40**, 1462–1475 (2001).
19. Voogt, J. A. & Oke, T. R. Thermal remote sensing of urban climates. *Remote Sens. Environ.* **86**, 370–384 (2003).
20. Shepherd, J. M. A review of current investigations of urban-induced rainfall and recommendations for the future. *Earth Interact.* **9**, 1–27 (2005).
21. Sharma, A. et al. Urban meteorological modeling using WRF: a sensitivity study. *Int. J. Climatol.* **37**, 1885–1900 (2017).
22. Salamanca, F., Martilli, A., Tewari, M. & Chen, F. A study of the urban boundary layer using different urban parameterizations and high-resolution urban canopy parameters with WRF. *J. Appl. Meteorol. Climatol.* **50**, 1107–1128 (2011).
23. Villarini, G. et al. Flood frequency analysis for nonstationary annual peak records in an urban drainage basin. *Adv. Water Resour.* **32**, 1255–1266 (2009).
24. Paciorek, C. J., Stone, D. A. & Wehner, M. F. Quantifying uncertainty in the attribution of human influence on severe weather. Preprint at <https://arxiv.org/abs/1706.03388> (2017).
25. Knutson, T. R. et al. Tropical cyclones and climate change. *Nat. Geosci.* **3**, 157–163 (2010).
26. Sobel, A. H. et al. Human influence on tropical cyclone intensity. *Science* **353**, 242–246 (2016).
27. Fang, Z., Dolan, G., Sebastian, A. & Bedient, P. B. Case study of flood mitigation and hazard management at the Texas Medical Center in the wake of tropical storm Allison in 2001. *Nat. Hazards Rev.* **15**, 05014001 (2014).
28. Pielke, R. A. et al. Land use/land cover changes and climate: modeling analysis and observational evidence. *WIREs Clim. Chang.* **2**, 828–850 (2011).
29. Lawrence, D. M. et al. The CCSM4 land simulation, 1850–2005: assessment of surface climate and new capabilities. *J. Clim.* **25**, 2240–2260 (2012).
30. Li, D., Malyshov, S. & Sheviakova, E. Exploring historical and future urban climate in the Earth System Modeling framework: 2. Impact of urban land use over the continental United States. *J. Adv. Model. Earth Syst.* **8**, 936–953 (2016).

Acknowledgements This material is based in part on work supported by the National Science Foundation under CAREER grant AGS-1349827 (to G.V.), NSF grant EAR-1520683 (to J.A.S. and G.A.V.), NSF grant AGS-1522492 and grant CBET-1444758 (to J.A.S.), and award NA14OAR4830101 from the National Oceanic and Atmospheric Administration, US Department of Commerce. G.A.V. was supported in part by The Carbon Mitigation Initiative at Princeton University.

Reviewer information *Nature* thanks A. Sharma and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions W.Z. and G.V. designed the experiments and performed the analyses. All authors interpreted the results and wrote the paper.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0676-z>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0676-z>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to G.V.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

The observed rainfall data are obtained from the Stage IV Quantitative Precipitation Estimates products over the continental USA (CONUS, <http://www.emc.ncep.noaa.gov/mmb/ylin/pcpanl/stage4/>) released by the National Centers for Environmental Prediction (NCEP). The discharge data are obtained from the United States Geological Survey (USGS). The gauged daily precipitation data are downloaded from the Global Historical Climatology Network (<ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/>).

The numerical simulations for hurricane Harvey are performed with the Advanced Research WRF (ARW) core of the Weather Research and Forecasting (WRF) model version 3.8. The land-use data used for the WRF experiments are the National Land Cover Database 2011 (NLCD2011)³¹ having 40 land-use categories including the types related to urban land-use: Developed Open Space (category 23), Developed Low Intensity (category 24), Developed Medium Intensity (category 25) and Developed High Intensity (category 26). In these four categories, the Developed Open Space category represents “areas with a mixture of some constructed materials, but mostly vegetation in the form of lawn grasses. Impervious surfaces account for less than 20% of total cover in this category”³². Because most of the surface of the Developed Open Space category is actually vegetation in the form of lawn grasses, we remapped only Developed Low Intensity (category 24), Developed Medium Intensity (category 25) and Developed High Intensity (category 26) into low-density residential, high-density residential and commercial use, respectively, by modifying the VEGPARM.TBL table under the WRF directory. We have plotted the remapped land-use map in the Houston area (Extended Data Fig. 7a). We run two sets of WRF experiments: one with the original land-use types by coupling Noah with the multi-layer building energy model (‘Urban BEM’)^{33–35} and the other by replacing the urban land-use types with croplands (‘NoUrban’)³⁶. The subtraction of precipitation in the multi-member ‘NoUrban’ from ‘Urban BEM’ experiments provides the impacts of urbanization on precipitation. ‘Urban BEM’ can be used to quantify the role of anthropogenic heat, given that previous studies have identified its influence on precipitation in urban areas^{37–40}. The WRF experiments (‘Urban BEM’ and ‘Urban BULK’) in this study are designed for hurricane Harvey, which is a very strong and large system for which the improvements from the BEM model over the ‘Urban BULK’ model may be muted; however, this may not be the case for more localized convective rain events, for which the differences due to the urban schemes may be more important. The Noah land surface model is used and coupled with WRF ARW in the ‘Urban BULK’ experiments. The simulations with the BEM model allow us to account for the direct anthropogenic heat emission because anthropogenic heat is directly computed. We use three domains for the WRF simulations in this study with spatial resolution of 12 km, 4 km and 1.33 km, respectively (Extended Data Fig. 7b). The inside domain d03 covers Houston while the outer domain d01 covers the region 104° W to 87° W and 24° N to 25° N. Feedback is allowed from the nesting domains to their parent domains. To account for uncertainty, we have performed seven experiments for both the ‘Urban BEM’ and ‘NoUrban’ experiments by initializing WRF ARW on 23 August 0 h to 24 August 12 h 2017 at 6-h intervals. The initial and boundary conditions for the WRF runs are 3-h NCEP North American Regional Reanalysis products⁴¹ (<https://rda.ucar.edu/datasets/ds608.0/>) on the Eta 221 grid at 29 pressure levels. The physics options are shown in Extended Data Table 1b.

The annual maximum flood peak data are based on the USGS measurements for the five watersheds shown in Extended Data Fig. 5 and Extended Data Table 1a (rather than using block maxima, an alternative approach would have been to use point processes, as has been advocated for in attribution studies⁴²). At each site and for each of the annual maxima we computed the basin averaged rainfall by spatially interpolating the available daily rain gauges from the Global Historical Climatology Network. To identify the rainfall accumulation that is more closely correlated with the flood peaks, we computed the Spearman correlation between the flood peaks and the accumulated rainfall from the day of the event up to a week before it; the rainfall accumulation window with the largest Spearman correlation coefficient with respect to the annual maximum peak discharge was used for the analyses. We use population (divided by 10^{-4} so that its range of values is comparable to that for precipitation) as a proxy for urbanization^{43–45}, and interpolate the information from the US Census Bureau for the City of Houston from the decadal surveys (<https://www.census.gov/prod/www/decennial.html>) with a third-order polynomial function to obtain annual estimates, and assuming that all the watersheds have experienced the same urban growth. For the development of the statistical models, we considered

two 2-parameter distributions (lognormal and gamma; see also refs^{46–48}), and made their parameters linearly depend (via appropriate link functions) on every combination of the two predictors, with the only constraint that the location parameter included at least precipitation. Model selection is performed via the Schwarz Bayesian criterion as a compromise between accuracy and parsimony. We assess the goodness of fit of the models by means of visual examination of the worm plots⁴⁹ (Extended Data Fig. 6) and by computing the first four moments of the residuals and their Filliben correlation coefficient (Extended Data Table 2).

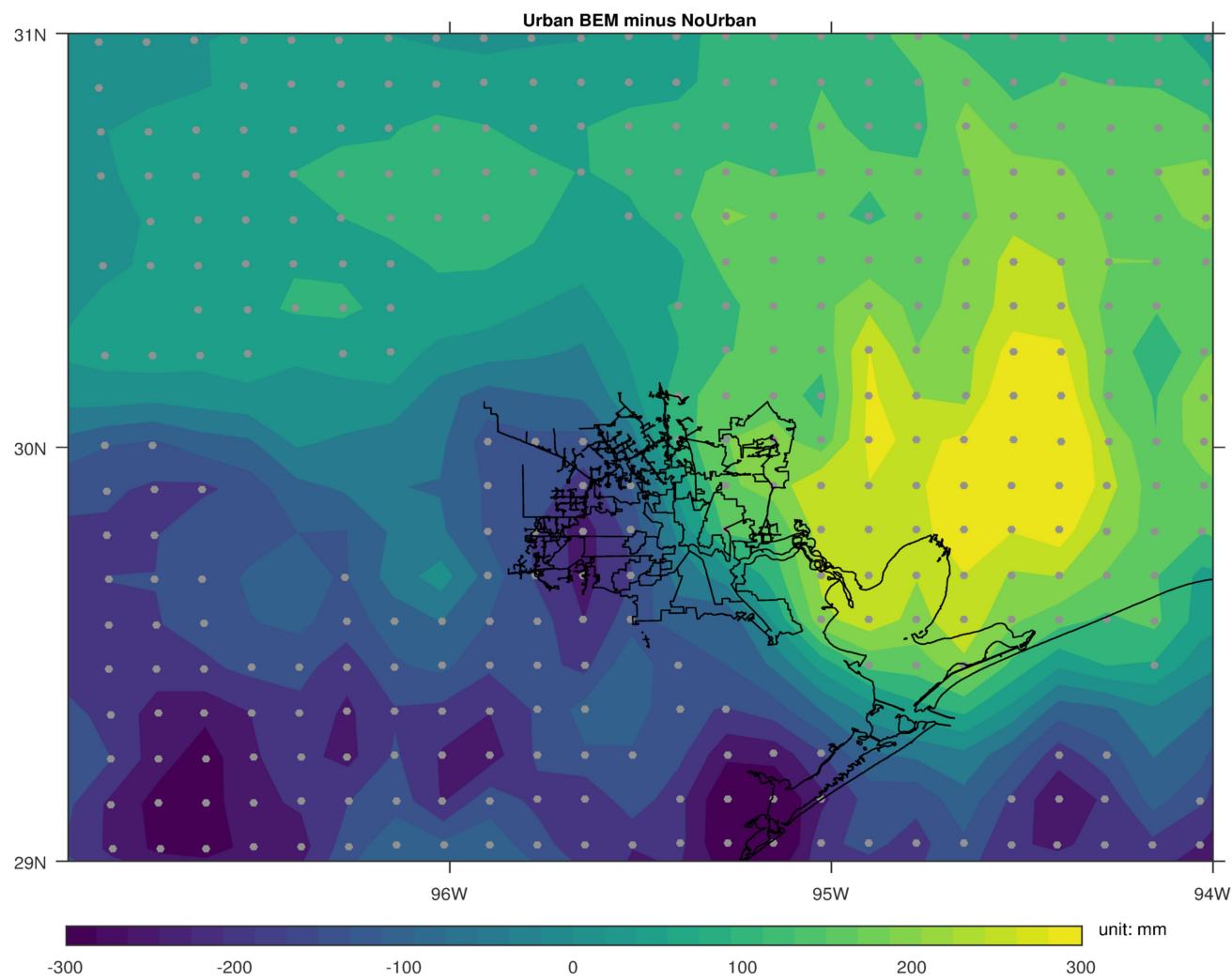
The risk ratio²⁴ represents the ratio between the probability of an event under the factual scenario (it is what we experienced and it is based on the observed urban intensity and rainfall) and the counterfactual scenario (it is what we would have experienced without urbanization).

Code availability. The codes related to the statistical modelling are available in Supplementary Information. The Advanced Research WRF (ARW) core of the WRF model version 3.8 was used to perform the simulations.

Data availability

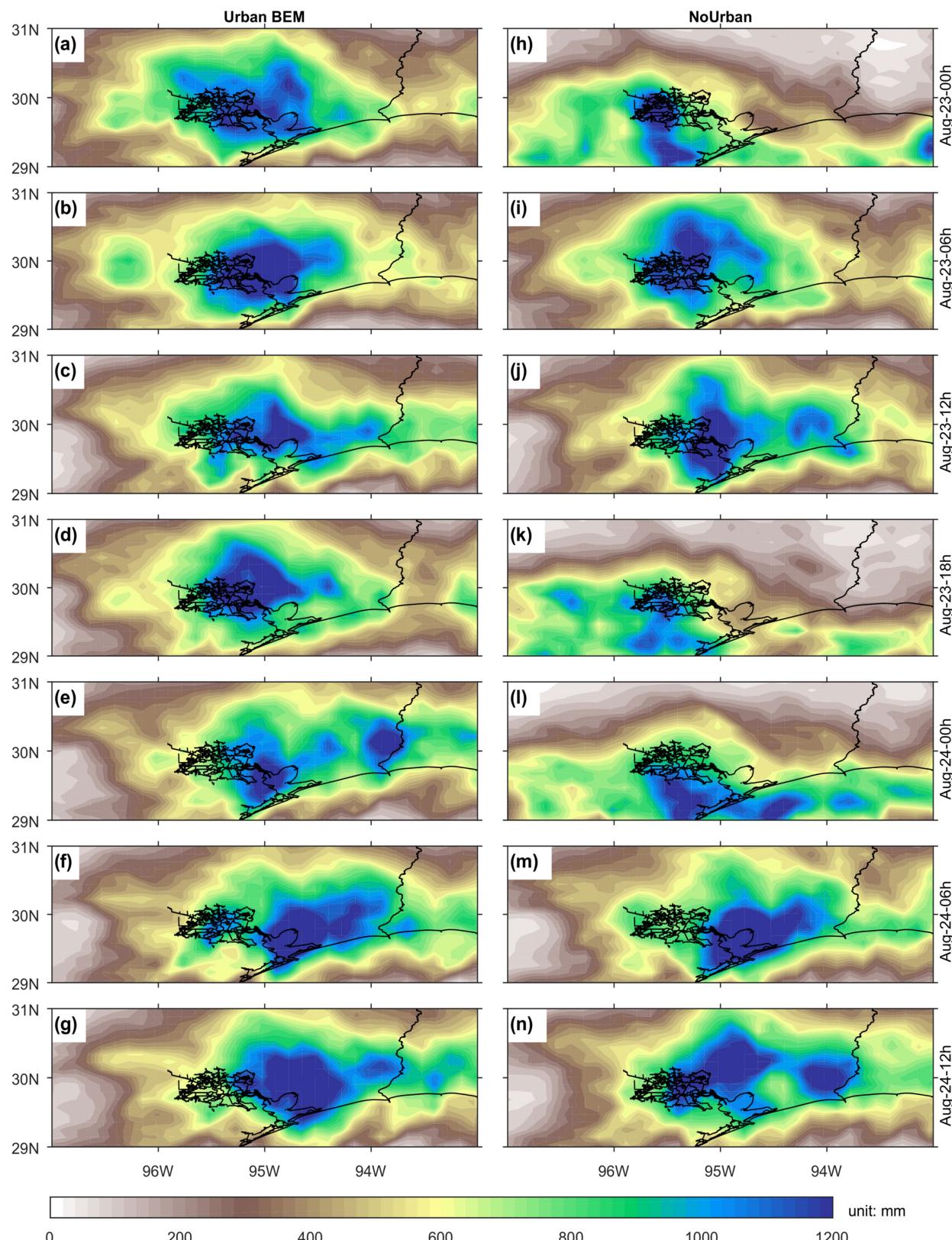
The data related to the statistical modelling are available in Supplementary Information. The additional data that support the findings of this study are available from the corresponding author upon reasonable request.

31. Jin, S. et al. A comprehensive change detection method for updating the National Land Cover Database to circa 2011. *Remote Sens. Environ.* **132**, 159–175 (2013).
32. Homer, C. et al. Completion of the 2011 National Land Cover Database for the conterminous United States—representing a decade of land cover change information. *Photogramm. Eng. Remote Sens.* **81**, 345–354 (2015).
33. Chen, F. et al. The integrated WRF/urban modelling system: development, evaluation, and applications to urban environmental problems. *Int. J. Climatol.* **31**, 273–288 (2011).
34. Li, D., Bou-Zeid, E., Baeck, M. L., Jessup, S. & Smith, J. A. Modeling land surface processes and heavy rainfall in urban environments: sensitivity to urban surface representations. *J. Hydrometeorol.* **14**, 1098–1118 (2013).
35. Lee, S.-H. et al. Evaluation of urban surface parameterizations in the WRF model using measurements during the Texas Air Quality Study 2006 field campaign. *Atmos. Chem. Phys.* **11**, 2127–2143 (2011).
36. Chen, F., Miao, S., Tewari, M., Bao, J. W. & Kusaka, H. A numerical study of interactions between surface forcing and sea breeze circulations and their effects on stagnation in the greater Houston area. *J. Geophys. Res.* **116**, D12105 (2011).
37. Kusaka, H., Nawata, K., Suzuki-Parker, A., Takane, Y. & Furuhashi, N. Mechanism of precipitation increase with urbanization in Tokyo as revealed by ensemble climate simulations. *J. Appl. Meteorol. Climatol.* **53**, 824–839 (2014).
38. Holst, C. C., Tam, C.-Y. & Chan, J. C. L. Sensitivity of urban rainfall to anthropogenic heat flux: a numerical experiment. *Geophys. Res. Lett.* **43**, 2240–2248 (2016).
39. Zhong, S. et al. Urbanization-induced urban heat island and aerosol effects on climate extremes in the Yangtze River Delta region of China. *Atmos. Chem. Phys.* **17**, 5439–5457 (2017).
40. Paul, S. et al. Increased spatial variability and intensification of extreme monsoon rainfall due to urbanization. *Sci. Rep.* **8**, 3918 (2018).
41. Mesinger, F. et al. North American regional reanalysis. *Bull. Am. Meteorol. Soc.* **87**, 343–360 (2006).
42. Prosdocimi, I., Kjeldsen, T. & Miller, J. Detection and attribution of urbanization effect on flood extremes using nonstationary flood-frequency models. *Wat. Resour. Res.* **51**, 4244–4262 (2015).
43. DeWalle, D. R., Swistock, B. R., Johnson, T. E. & McGuire, K. J. Potential effects of climate change and urbanization on mean annual streamflow in the United States. *Wat. Resour. Res.* **36**, 2655–2664 (2000).
44. Gluck, W. R. & McCuen, R. H. Estimating land use characteristics for hydrologic models. *Wat. Resour. Res.* **11**, 177–179 (1975).
45. Stankowski, S. J. Population density as an indirect indicator of urban and suburban land-surface modifications. *US Geol. Surv. Prof. Pap.* **800**, 219–224 (1972).
46. Villarini, G., Serinaldi, F., Smith, J. A. & Krajewski, W. F. On the stationarity of annual flood peaks in the continental United States during the 20th century. *Wat. Resour. Res.* **45**, W08417 (2009).
47. López, J. & Francés, F. Non-stationary flood frequency analysis in continental Spanish rivers, using climate and reservoir indices as external covariates. *Hydrol. Earth Syst. Sci.* **17**, 3189–3203 (2013).
48. Slater, L. J. & Villarini, G. Evaluating the drivers of seasonal streamflow in the US Midwest. *Wat. Resour. Res.* **9**, 695 (2017).
49. van Buuren, S. & Fredriks, M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Stat. Med.* **20**, 1259–1277 (2001).



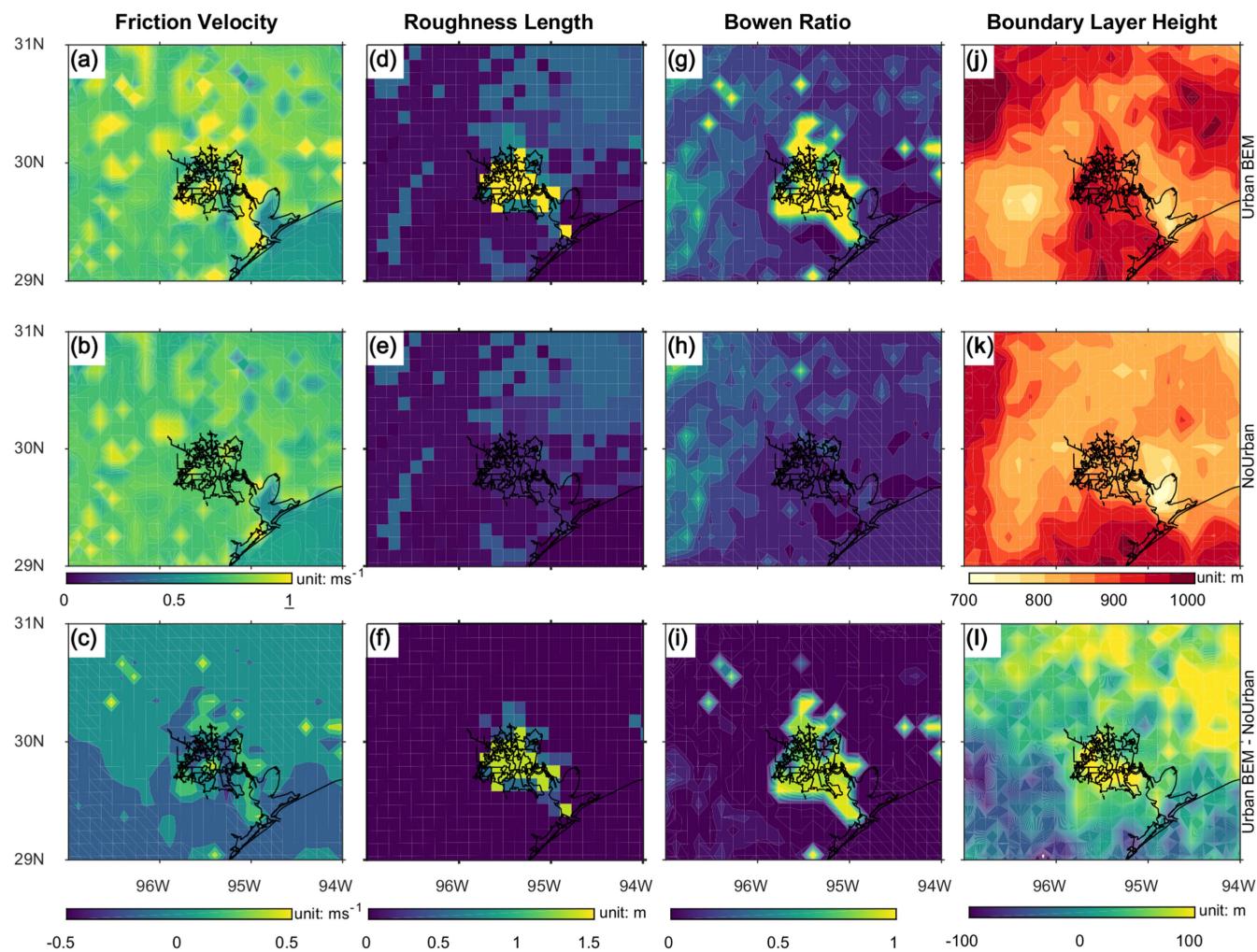
Extended Data Fig. 1 | Effect of urbanization on the storm total rainfall during hurricane Harvey. The map (the black outlines mark urban areas of Houston) shows the difference (Urban BEM minus NoUrban) in accumulated precipitation for 25 August 0 h to 30 August 0 h 2017

between the 'Urban BEM' and 'NoUrban' WRF experiments. The stippled regions represent areas for which these differences are statistically different from zero (that is, there are no effects of urbanization in terms of rainfall) at the $P=0.05$ significance level based on Student's t test.



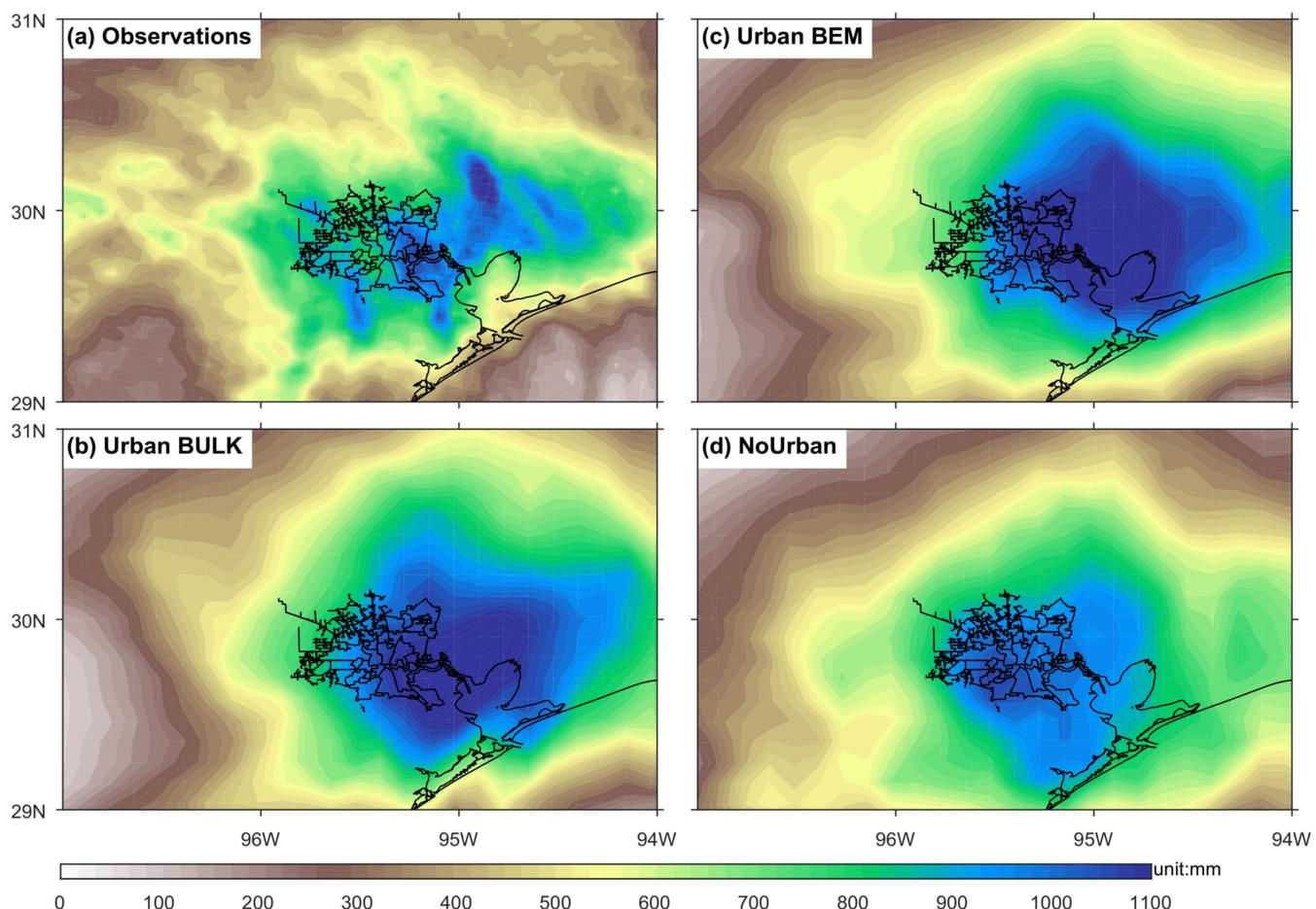
Extended Data Fig. 2 | Accumulated precipitation in each ensemble member of the WRF experiments. a–n, Accumulated precipitation (colour scale) for 25 August 0 h to 30 August 0 h 2017 in each member of

the 'Urban BEM' (a–g) and 'NoUrban' (h–n) WRF experiments initialized between 23 August 0 h and 24 August 12 h at 6-h intervals.



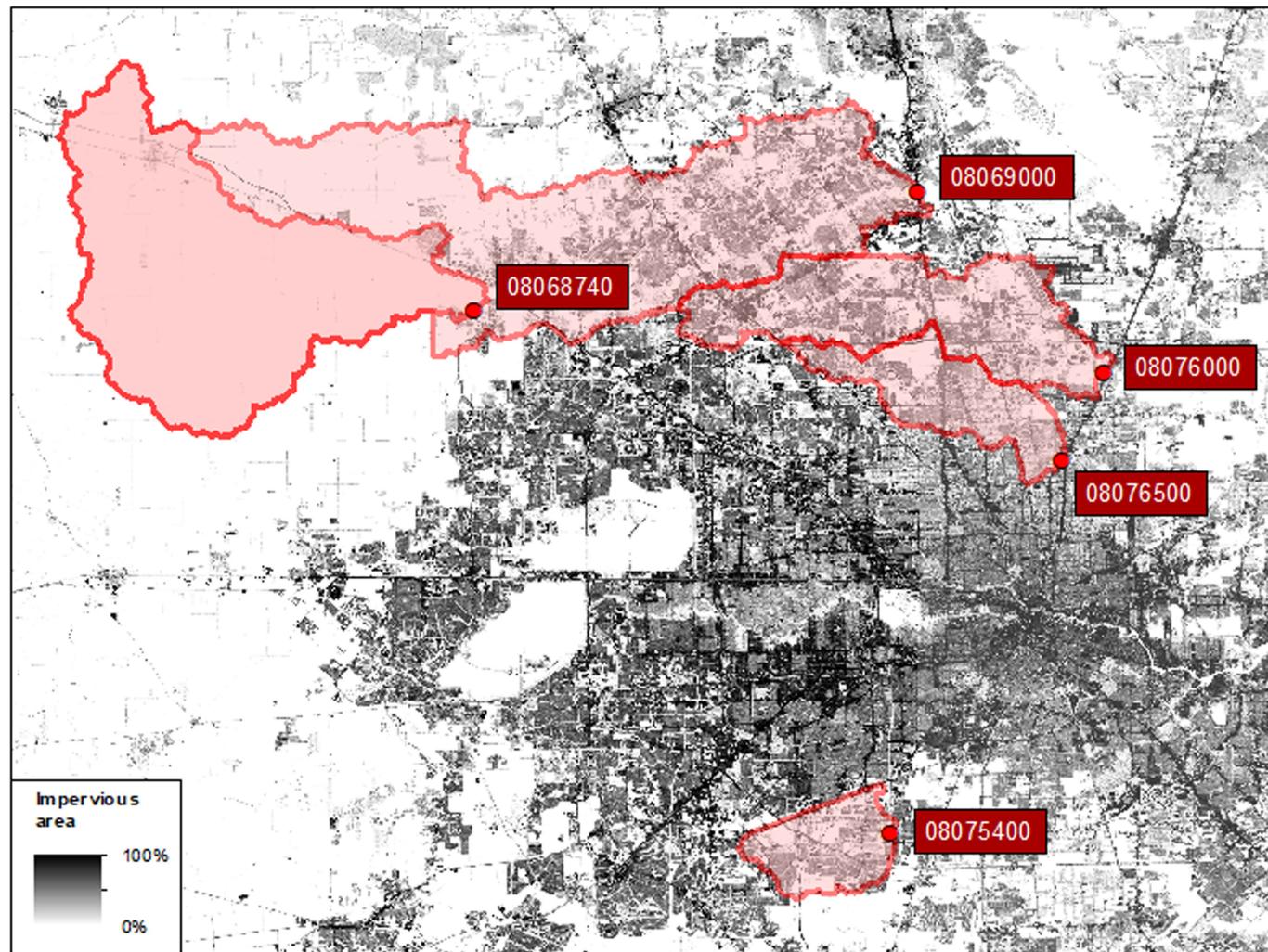
Extended Data Fig. 3 | Key variables for diagnosing the impacts of urbanization on hurricane Harvey. a–l, Friction velocity (a–c), roughness length (d–f), Bowen ratio (g–i) and boundary layer height (j–l) are

shown for the ‘Urban BEM’ (top panels) and ‘NoUrban’ (middle panels) experiments with WRF and their differences (bottom panels).

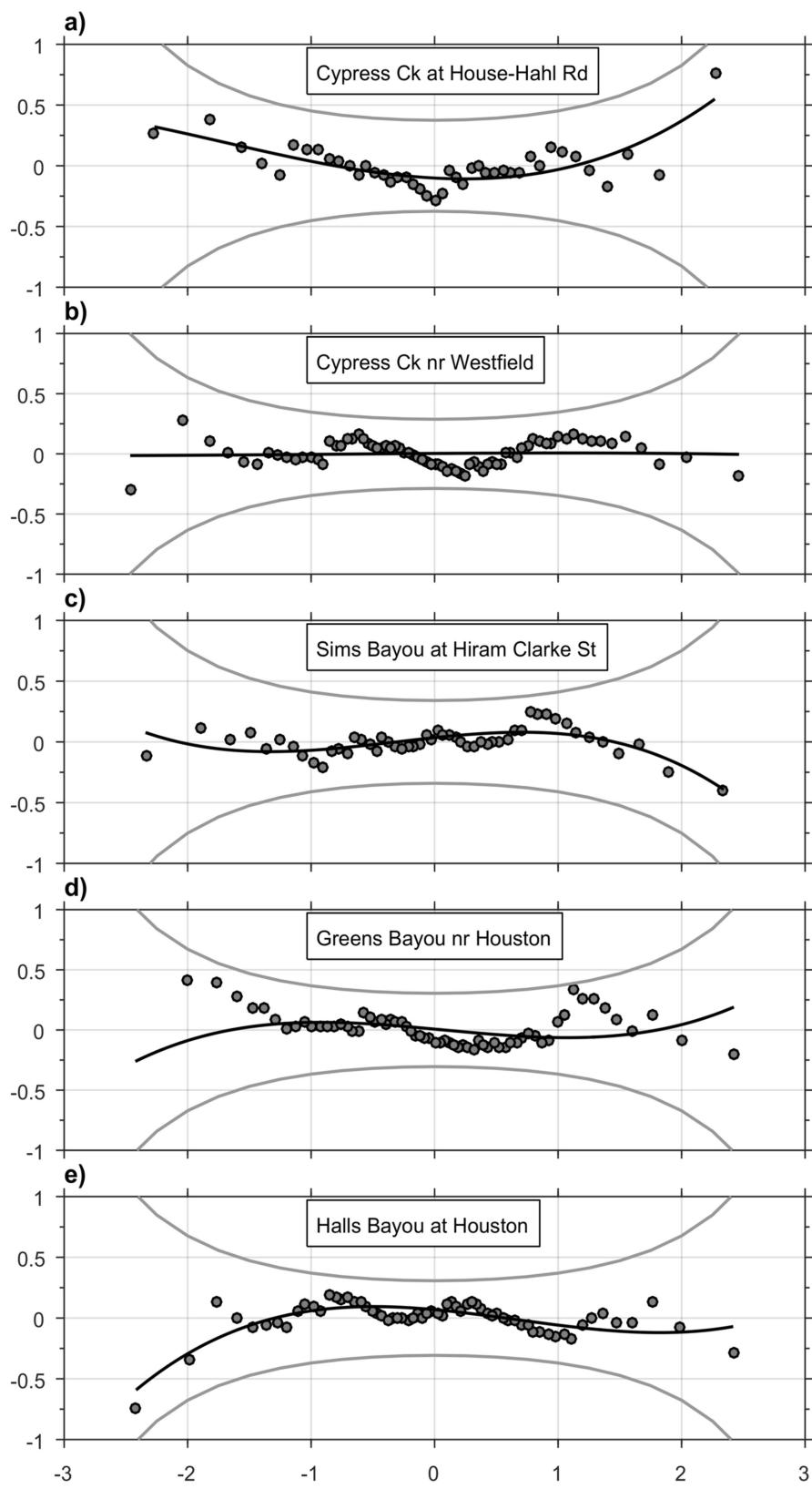


Extended Data Fig. 4 | Accumulated precipitation for hurricane Harvey in observations and different urbanization schemes and settings of WRF experiments. a–d, Accumulated precipitation (colour scale) is

shown for 25 August 0 h to 30 August 0 h 2017 in observations (a), and in the ‘Urban BULK’ (b), ‘Urban BEM’ (c) and ‘NoUrban’ (in which urban land-use types are replaced by croplands; d) WRF experiments.

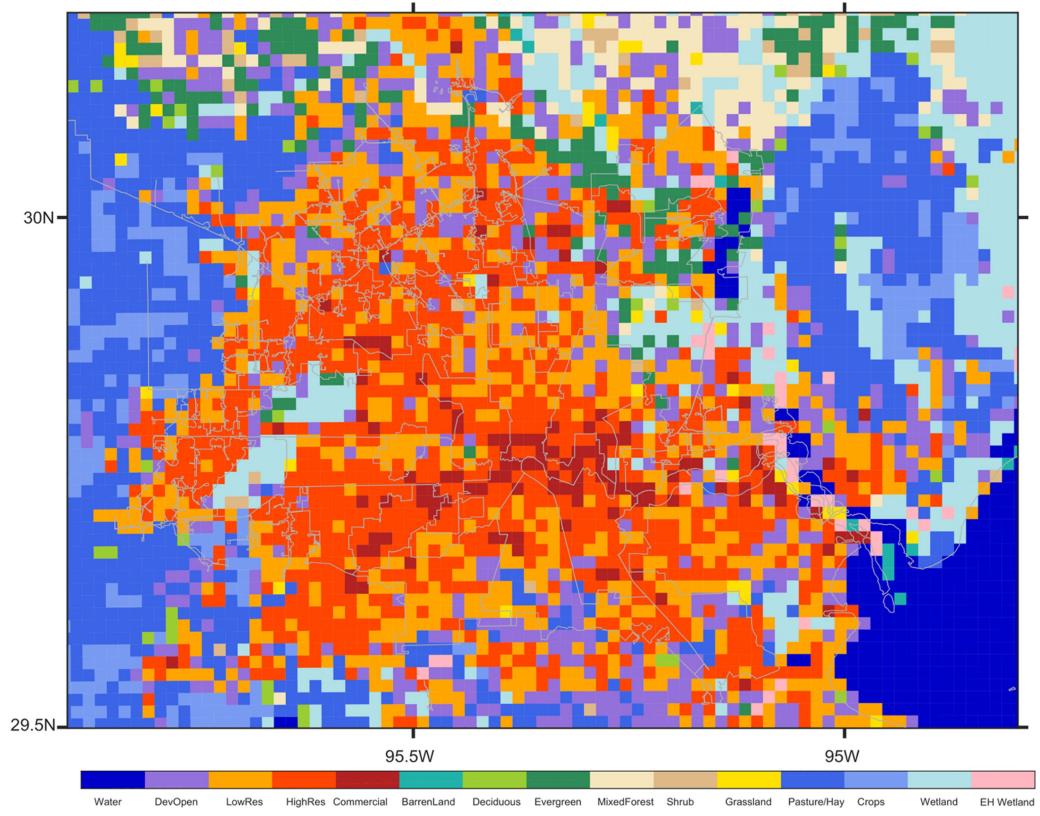


Extended Data Fig. 5 | Basin boundaries of the five watersheds considered in this study. The ID number for each basin is also shown. The percentage of impervious area is indicated by the grey scale.

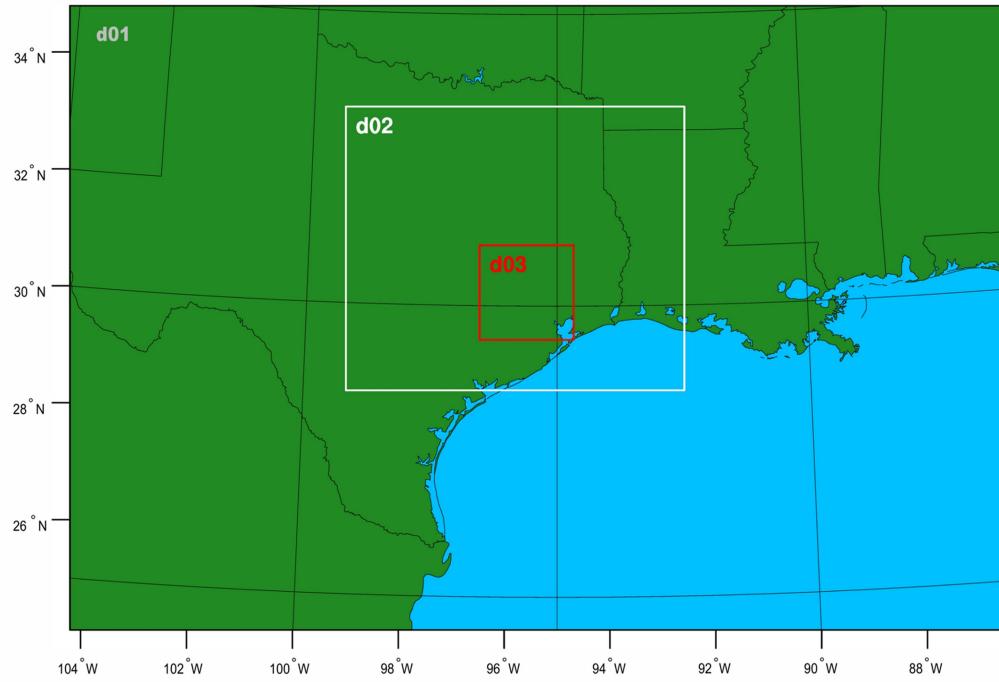


Extended Data Fig. 6 | Worm plots for the fitted models of annual maximum peak discharge records. a–e, Worm plots for the fitted models shown to evaluate the goodness of fit as shown in Fig. 3. For a satisfactory fit, the data points should be within the two grey lines (95% confidence interval).

a)



b)


Extended Data Fig. 7 | Information related to the WRF simulations.

a, Land-use map in the Houston area. The low-residential, high-residential and commercial land-use categories are coloured in orange, red and dark red, respectively. (DevOpen, developed open space; EH Wetland, emergent

herbaceous wetlands.) **b**, Three spatial domains d01, d02 and d03 in the WRF simulations with spatial resolution of 12 km, 4 km and 1.33 km, respectively.

Extended Data Table 1 | Summary of the characteristics of the five watersheds studied and of the WRF physics options

a)

USGS ID	Name	Drainage area (mi ²)	Risk Ratio
08068740	Cypress Creek at House-Hahl Road near Cypress, TX	131	1.1
08069000	Cypress Creek near Westfield, TX	285	1.9
08075400	Sims Bayou at Hiram Clarke Street, Houston, TX	20.2	8.3
08076000	Greens Bayou near Houston, TX	68.7	91.9
08076500	Halls Bayou at Houston, TX	28.7	2.9

b)

Physics	Options
Microphysics	WSM 6-class graupel scheme
Surface layer	Monin-Obukhov scheme
Land surface	unified Noah land-surface model
Boundary layer scheme	Mellor-Yamada-Janjic TKE scheme
Cumulus parameterization	None for d02 and d03, and the Betts-Miller-Janjic scheme for d01
Longwave radiation	Rapid Radiative Transfer Model
Shortwave radiation	Dudhia scheme
Land use	NLCD2011 (40 categories)

a. Summary information about the five basins considered in this study, and the related value of the risk ratio (see Methods). **b.** Setting of WRF physics options. (WSM, the WRF Single-moment Microphysics scheme.)

Extended Data Table 2 | Summary of the modelling results for the five basins considered in this study

	08068740	08069000	08075400	08076000	08076500
Distribution	Lognormal	Gamma	Gamma	Lognormal	Gamma
Intercept (μ)	6.96 (0.12)	7.42 (0.21)	7.04 (0.35)	7.08 (0.20)	7.21 (0.17)
Rainfall (μ)	0.0068 (0.001)	0.0026 (0.0004)	0.0032 (0.001)	0.0044 (0.001)	0.003 (0.001)
Population (μ)	-	0.0051 (0.001)	0.0055 (0.002)	0.0070 (0.001)	0.003 (0.001)
Intercept (σ)	-0.51 (0.11)	-0.11 (0.20)	-0.65 (0.09)	-0.75 (0.09)	-0.94 (0.08)
Rainfall (σ)	-	-	-	-	-
Population (σ)	-	-0.004 (0.001)	-	-	-
Mean (residuals)	0.00	0.00	0.00	0.00	0.00
Variance (residuals)	1.02	1.02	1.02	1.02	1.02
Skewness (residuals)	0.55	-0.02	-0.14	-0.17	-0.41
Kurtosis (residuals)	3.27	2.77	2.29	3.83	3.51
Filliben (residuals)	0.984	0.995	0.993	0.982	0.990

The first value is the point estimate, and the value in parentheses is the standard error. The '-' symbol means that the parameter does not depend on the predictor. For the lognormal distribution the link function for the μ parameter is the identity, whereas for the σ parameter the link function is logarithmic. For the gamma distribution both parameters have a logarithmic link function.

Water input into the Mariana subduction zone estimated from ocean-bottom seismic data

Chen Cai^{1*}, Douglas A. Wiens¹, Weisen Shen^{1,2} & Melody Eimer¹

The water cycle at subduction zones remains poorly understood, although subduction is the only mechanism for water transport deep into Earth. Previous estimates of water flux^{1–3} exhibit large variations in the amount of water that is subducted deeper than 100 kilometres. The main source of uncertainty in these calculations is the initial water content of the subducting uppermost mantle. Previous active-source seismic studies suggest that the subducting slab may be pervasively hydrated in the plate-bending region near the oceanic trench^{4–7}. However, these studies do not constrain the depth extent of hydration and most investigate young incoming plates, leaving subduction-zone water budgets for old subducting plates uncertain. Here we present seismic images of the crust and uppermost mantle around the central Mariana trench derived from Rayleigh-wave analysis of broadband ocean-bottom seismic data. These images show that the low mantle velocities that result from mantle hydration extend roughly 24 kilometres beneath the Moho discontinuity. Combined with estimates of subducting crustal water, these results indicate that at least 4.3 times more water subducts than previously calculated for this region³. If other old, cold subducting slabs contain correspondingly thick layers of hydrous mantle, as suggested by the similarity of incoming plate faulting across old, cold subducting slabs, then estimates of the global water flux into the mantle at depths greater than 100 kilometres must be increased by a factor of about three compared to previous estimates³.

Because a long-term net influx of water to the deep interior of Earth is inconsistent with the geological record⁸, estimates of water expelled at volcanic arcs and backarc basins probably also need to be revised upwards⁹.

The Mariana subduction zone has long been cited as a water-rich system owing to the prevalence of forearc serpentinite mud volcanoes¹⁰, a serpentized mantle wedge¹¹, and hydrous arc and backarc lavas^{12,13}. However, the initial amount of water within the subducting mantle is unknown. The subducting Pacific plate in this region is among the oldest sections of oceanic lithosphere worldwide¹⁴ (about 150 Myr). The incoming plate displays widespread plate-bending normal-fault scarps and earthquakes^{15,16} (Fig. 1b), making it an excellent place to investigate the depth extent of faulting-induced hydration in an old oceanic plate. The potential hydration of old, cold oceanic plates is particularly important for the water cycle because the thermal structure of the plates permits temperature-sensitive hydrous minerals to occur throughout a thicker region².

We used data collected by broadband ocean-bottom seismographs and island seismographs deployed around the central Mariana trench (Fig. 1a). The deployment covers a sufficiently large region to investigate structure variations in the subducting plate before and after subduction. We derived local Rayleigh-wave group- and phase-velocity dispersion curves using two surface-wave methods (Methods). A Bayesian Monte Carlo algorithm was then used to determine a

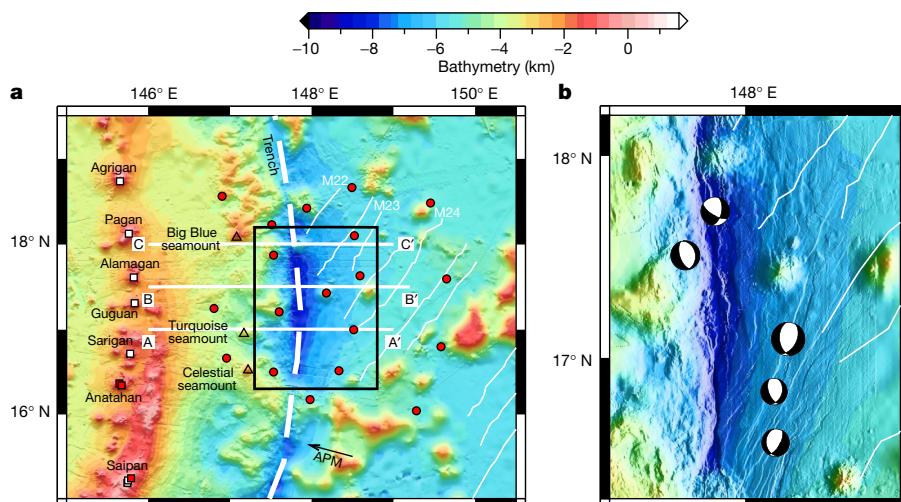


Fig. 1 | Distribution of seismic stations and bathymetry. **a**, Station distribution. Red circles show ocean-bottom seismographs deployed from January 2012 to February 2013. White squares represent the temporary island-based stations. Red squares indicate stations from the US Geological Survey (USGS) Northern Mariana Islands Seismograph Network used in our study. Open triangles show the locations of three large serpentine seamounts within the study area. The dashed white line is the trench axis. Thick solid white lines show the cross-section

locations in Fig. 2a–c. The arrow labelled APM indicates the direction of absolute plate motion. Thin solid white lines show magnetic lineations (M22, M23, M24)³⁰. **b**, High-resolution bathymetry for the outer-rise region of the Mariana trench (indicated by the black rectangle in **a**) and relocated earthquake locations from the Global Centroid Moment Tensor catalogue¹⁵ (black and white focal projections). Magnetic lineations are shown as in **a**.

¹Department of Earth and Planetary Sciences, Washington University in St Louis, St Louis, MO, USA. ²Department of Geosciences, Stony Brook University, Stony Brook, NY, USA.
*e-mail: cai.chen@wustl.edu

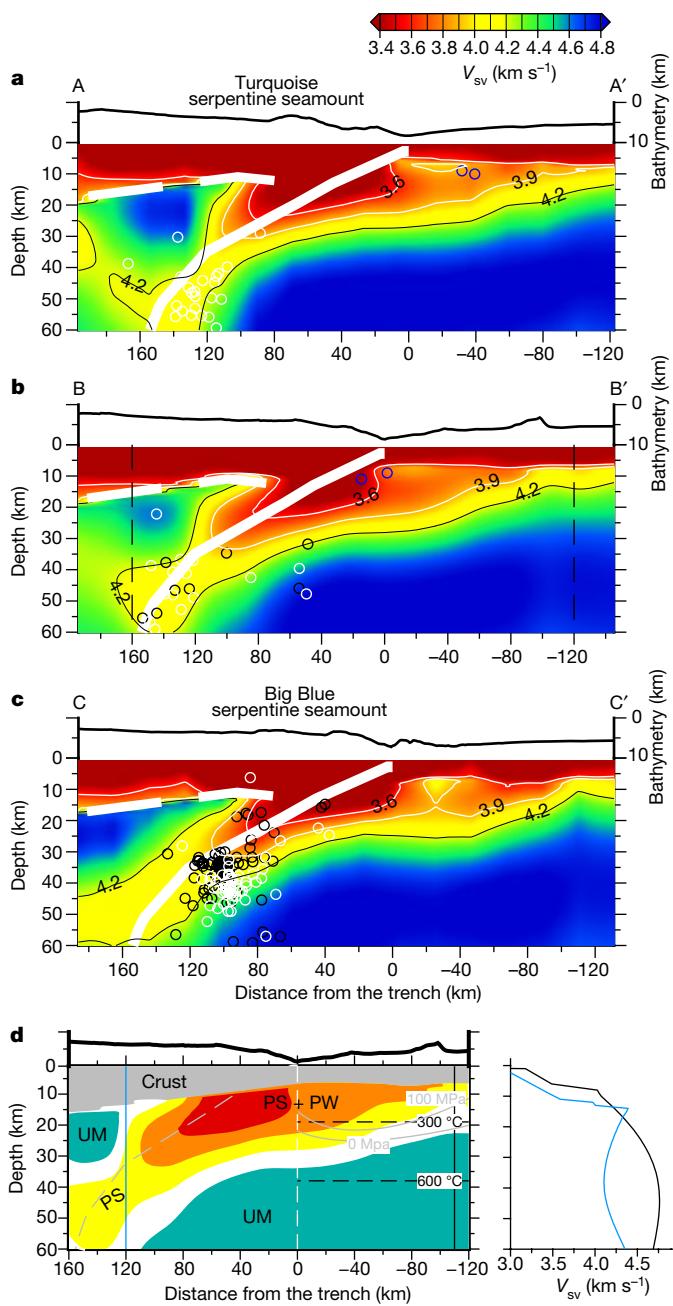


Fig. 2 | Vertical profiles and interpretation. **a–c**, Cross-sections A–A' (a), B–B' (b) and C–C' (c) showing the azimuthally averaged velocity of vertically polarized S waves (V_{sv} ; colour scale). Thick dashed white lines show the location of the forearc Moho³¹. Thick solid white lines are projected 6-km-thick slab crust (combining an active-source reflection result¹⁶ for depths of less than 30 km and the Slab 1.0 model³² for greater depths). Thin white lines are velocity contours of 3.6 km s^{-1} and 3.9 km s^{-1} , and thin black lines are velocity contours of 4.2 km s^{-1} . Black³³, white¹¹ and blue¹⁵ circles are relocated earthquakes in the subducting plate around each profile from previous studies. **d**, Interpretation of seismic structure in b within the area bounded by the two vertical dashed black lines. Dashed black lines are isotherms of 300°C and 600°C calculated from ref. ²⁵. Solid grey lines are contours of 0 MPa and 100 MPa extensional stress¹⁵. The dashed grey line marks the surface of the slab. The vertical dashed white line indicates the trench axis. Vertical solid blue and black lines mark the locations of the one-dimensional V_{sv} structure shown on the right (colour-coded). The colouring schematically illustrates the velocity, as in a–c. UM, unaltered mantle; PS, partial serpentinization; PW, pore water.

three-dimensional image of the shear-wave velocity of the crust and uppermost mantle (Methods). Compared to previous active-source seismic studies of other subduction zones^{4–7}, our study resolves the

three-dimensional structure to greater depth and avoids biases caused by possible azimuthal anisotropy. In addition, it images shear velocities that are more sensitive to hydration than are compressive velocities¹⁷.

The resulting azimuthally averaged velocity model of vertically polarized S waves shows systematic changes in the incoming plate and subducting slab along the direction normal to the trench (Fig. 2a–c). At distances far from the trench axis (more than 100 km seaward), the subducting plate has a typical structure of old oceanic lithosphere¹⁸ at depths of more than 20 km, with velocities greater than 4.5 km s^{-1} . A lower-velocity (about 4.2 km s^{-1}) layer with a thickness of about 10 km is observed immediately beneath the Moho, consistent with active-source seismic results¹⁹. This suggests that the very shallow part of the incoming plate starts to be altered far from the trench, as observed in other subduction zones⁴. A thicker region of low velocities begins about 80 km seaward of the trench axis and deepens towards the trench, with velocities as low as 3.8 km s^{-1} . At the trench axis, the bottom of this low-velocity layer reaches 30 ± 5 km beneath the seafloor, 24 ± 5 km into the upper mantle. (The errors quoted here and elsewhere correspond to the variations in thickness that result from different assumptions about the velocity that bounds the serpentine layer.) This low-velocity layer persists after the Pacific plate is subducted as a (30 ± 5) -km-thick low-velocity layer atop the fast slab mantle between 100 km and 160 km west of the trench axis. Synthetic data tests show that a velocity anomaly of this large magnitude and extent cannot result from the effect of limited resolution of an approximately 6-km-thick layer of low-velocity oceanic crust (Methods, Extended Data Fig. 1).

We also solved for the azimuthal anisotropy of the phase velocity. These results show trench-parallel fast axes in the incoming Pacific plate for periods of 12–21 s, with a maximum magnitude as large as 9% reached between 14 s and 16 s. By contrast, for periods of more than 27 s, the fast directions rotate to be oblique to the trench strike, close to the palaeo-spreading direction, which is normal to the magnetic lineations (Fig. 3).

The region of the incoming plate where we observe a reduction in mantle velocity and large azimuthal anisotropy coincides with the plate-bending region, as characterized by substantial normal faulting seismicity¹⁵ and large extensional seafloor fault scarps¹⁶ (Fig. 1b). There is a clear spatial association between plate-bending-induced faulting and velocity reduction. Velocities within the incoming plate begin to decrease sharply 80 km from the trench, at about the same distance at which intense seismicity and faulting begins on the seafloor¹⁶. In this region, the fault scarps and the earthquake fault planes strike approximately north–south, subparallel to the trench axis (Fig. 1b). This is consistent with the observations of the azimuthal anisotropy of the phase velocity at 12–21 s, which primarily sample depths down to about 25 km below the seafloor (Fig. 3). These results reveal trench-parallel fast directions, as would be expected for pervasive trench-parallel water-filled faults or zones of alteration. The flexure model that best fits the bathymetry of the Pacific plate seaward of the trench axis¹⁵ predicts a neutral plane at around 30 km (Fig. 2d), which suggests that brittle normal faulting can extend nearly 30 km into the plate. This prediction agrees well with the maximum depth extent (30 ± 5 km) of the low-velocity zone that we observed in the incoming plate near the trench in our study (Fig. 2a–c), which suggests that this zone is related directly to brittle normal faults.

Many previous studies at other subduction zones have attributed low-velocity zones in the upper mantle associated with plate bending to the hydration of mantle peridotite to form low-velocity serpentine minerals. Extensional deformation within the shallow part of the incoming plate produces a pressure gradient that may enable water to penetrate deep into the slab along normal faults²⁰. The serpentinization rate is geologically fast if water delivery to the serpentinization front is efficient^{21,22}. Alternatively, other studies attribute the reductions in mantle velocity to water-filled porosity and cracks²³. However, the potential velocity effect of water-filled cracks is usually difficult to estimate directly because it depends critically on the aspect ratio and

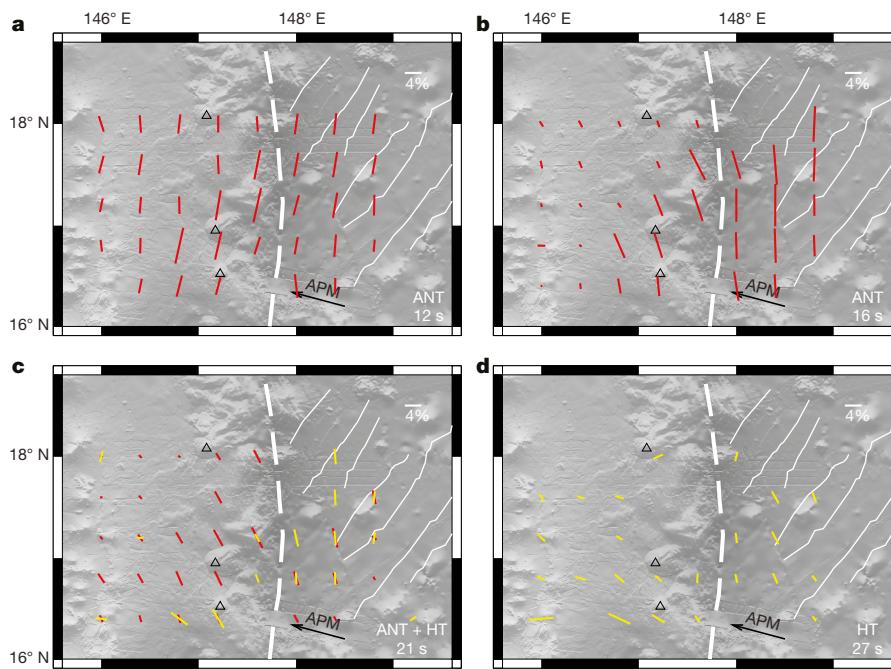


Fig. 3 | Azimuthal anisotropy results at various periods. **a, b,** Results from ambient noise tomography (ANT; red bars) at periods of 12 s (**a**) and 16 s (**b**). **c,** Results from ambient noise tomography and Helmholtz tomography (HT; yellow bars) at a period of 21 s. **d,** Results from Helmholtz tomography at a period of 27 s. The orientations of the red

and yellow bars represent the fast directions and the lengths indicate the magnitude of anisotropy. The trench axis, serpentine seamounts, absolute plate motion and magnetic lineations are shown and labelled as in Fig. 1a. Only nodes with good azimuthal coverage and good azimuthal anisotropy fittings are plotted.

spatial density of the cracks, which are largely unknown. In this study, we use the increase in velocity as the plate subducts and porosity is reduced to distinguish the effects of water-filled cracks, porosity and serpentinization (Fig. 2d).

The seismic images of the Mariana trench show that the velocity of the low-velocity zone increases as the top of the slab subducts past a depth of about 30 km (Fig. 2a–c). Reductions in porosity due to increased pressure²⁴ reduce or eliminate the velocity effect of water-filled cracks at depth, whereas hydrous minerals remain stable at the cold slab temperatures; this provides a means to separate the complementary velocity effects of porosity and altered minerals and to estimate the concentration of hydrous minerals (PS region in Fig. 2d). Compared to the low-velocity zone within the slab before subduction, the low-velocity zone within the subducted-slab mantle at depths of around 40 km preserves the original thickness (30 ± 5 km) but exhibits a smaller reduction in velocity (about 4.1 km s^{-1} ; Fig. 2a–c). We use this smaller velocity reduction to estimate the degree of mantle serpentinization in the downgoing slab. Therefore, we base our estimates of the water content due to serpentinization of the subducting slab on the initial thickness of the low-velocity layer at the trench and on the shear-wave velocities of around 4.1 km s^{-1} observed at depths of 30–50 km in the subducting plate mantle, after most of the pore water is expelled (PS region in Fig. 2d). The additional velocity reduction (about 0.3 km s^{-1}) within the low-velocity zone in the slab before subduction can be attributed to pore water in cracks (PS + PW region in Fig. 2d).

Calibrating the change in seismic velocity to the degree of serpentinization requires knowledge of the seismic velocity of serpentine. We select lizardite, the form of serpentine expected to predominate at lower temperatures²², to interpret the observed velocity reduction. The nominal temperature predicted by plate cooling models²⁵ around 30 km beneath the seafloor is roughly 470°C (Fig. 2d), possibly higher than the temperature of lizardite breakdown (about 320°C in ref. ²⁶, although lizardite has been found at temperatures as high as 580°C in ref. ²²). Water circulation in cracks may lower the temperature of the slab mantle in the plate-bending region into the lizardite stability field. In addition, selecting lizardite provides an estimate of the lower bound

of water input (Methods). Using the experimental relationship between the change in shear velocity and the serpentine volume fraction for lizardite²⁷, the shear velocity of 4.1 km s^{-1} observed within the subducted slab corresponds to a change in velocity of about 0.41 km s^{-1} (Methods), indicating roughly 19 vol% serpentinization (about 2 wt% water). It is possible that the anisotropic effects of serpentine distributed along bending faults could cause additional reductions in velocity, leading to an overestimate of the serpentinization percentage²⁸. However, according to our calculation with a realistic dipping-fault geometry and the frequencies used in this study, the anisotropic effects are not important when estimating the serpentinization percentage (Methods, Extended Data Fig. 2).

We therefore interpret the seismic images as strong evidence for a (24 ± 5) -km-thick, partially serpentinized (2 wt% water) slab-mantle layer. Applying a convergence rate of 50 mm yr^{-1} , the amount of water input into the Mariana subduction zone through mantle serpentinization would be about $79 \pm 17 \text{ Tg Myr}^{-1} \text{ m}^{-1}$; the total water flux is approximately $94 \pm 17 \text{ Tg Myr}^{-1} \text{ m}^{-1}$ if water in the sediment and crust is also included from previous estimates³. This estimate of water flux into the Mariana trench is 4.3 ± 0.8 times larger than a previous estimate³, which assumed a 2-km-thick, partially serpentinized slab mantle (2 wt% water). All uncertainties are estimated on the basis of the uncertainty of the thickness of the serpentinized slab mantle.

Our interpretation of serpentinization extending to depths of around 24 km below the Moho in the incoming plate at the Mariana trench has important implications for water flux into subduction zones globally. This depth is greater than the maximum observed depth of large normal-faulting earthquakes and close to the estimated depth of the neutral stress plane¹⁵ (Fig. 2d). The maximum depth of serpentinization near trenches has not been well determined for other older incoming plates, because the depth extent is too great to be well constrained by active-source seismic studies^{4–7} and surface-wave investigations have not been performed elsewhere. The bending and faulting features of the incoming Pacific plate near the Mariana trench are similar to those observed at other old subduction plates, and the maximum depth of normal faulting and the depth of the neutral plane are generally about the same²⁹. Therefore, it is reasonable that serpentinization

extends to similar depths of 20–25 km below the Moho at other sites where old lithosphere subducts. Modifying previous calculations of global water flux into subduction zones to take into account the hydrous alteration of this increased mantle thickness yields an estimated flux of 3.0×10^9 Tg Myr $^{-1}$ (Methods)—an increase by a factor of about three^{1–3}.

This larger estimate of the input water flux at subduction zones is much greater than current estimates of water output from the mantle. Because a large long-term net influx of water to the deep interior is inconsistent with the stability of sea level in the geological record^{2,8}, one possible implication of our result is that the thick layer of serpentinized mantle that we find in the Mariana subduction zone is not characteristic of other old, cold subducting slabs, and that the Mariana slab carries much more water than other subduction zones. However, there is little indication that the incoming plate-bending region of the Mariana subduction zone is substantially different in terms of morphology and intensity of faulting compared to the corresponding regions of other old subduction zones. Thus, the most likely interpretation is that previous estimates of water output from the mantle are also underestimated. Estimates of water output from the mantle at mid-ocean ridges and ocean islands may be relatively well constrained, but estimates for volcanic arcs and backarcs rely on the melt flux and the water content, which are poorly constrained^{8,9}.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0655-4>.

Received: 17 April 2018; Accepted: 19 September 2018;

Published online 14 November 2018.

- Hacker, B. R. H₂O subduction beyond arcs. *Geochem. Geophys. Geosyst.* **9**, Q03001 (2008).
- Rüpke, L. H., Morgan, J. P., Hort, M. & Connolly, J. A. D. Serpentine and the subduction zone water cycle. *Earth Planet. Sci. Lett.* **223**, 17–34 (2004).
- van Keken, P. E., Hacker, B. R., Syracuse, E. M. & Abers, G. A. Subduction factory: 4. Depth-dependent flux of H₂O from subducting slabs worldwide. *J. Geophys. Res.* **116**, B01401 (2011).
- Fujie, G. et al. Systematic changes in the incoming plate structure at the Kuril trench. *Geophys. Res. Lett.* **40**, 88–93 (2013).
- Ranero, C. R., Phipps Morgan, J., McIntosh, K. & Reichert, C. Bending-related faulting and mantle serpentinization at the Middle America trench. *Nature* **425**, 367–373 (2003).
- Shillington, D. J. et al. Link between plate fabric, hydration and subduction zone seismicity in Alaska. *Nat. Geosci.* **8**, 961–964 (2015).
- Van Avendonk, H. J. A., Holbrook, W. S., Lizarralde, D. & Denyer, P. Structure and serpentinization of the subducting Cocos plate offshore Nicaragua and Costa Rica. *Geochem. Geophys. Geosyst.* **12**, Q06009 (2011).
- Parai, R. & Mukhopadhyay, S. How large is the subducted water flux? New constraints on mantle regassing rates. *Earth Planet. Sci. Lett.* **317–318**, 396–406 (2012).
- Grove, T. L., Till, C. B. & Krawczynski, M. J. The role of H₂O in subduction zone magmatism. *Annu. Rev. Earth Planet. Sci.* **40**, 413–439 (2012).
- Fryer, P. Serpentine mud volcanism: observations, processes, and implications. *Annu. Rev. Mar. Sci.* **4**, 345–373 (2012).
- Barklage, M. et al. P and S velocity tomography of the Mariana subduction system from a combined land-sea seismic deployment. *Geochem. Geophys. Geosyst.* **16**, 681–704 (2015).
- Kelley, K. A. et al. Mantle melting as a function of water content beneath the Mariana arc. *J. Petrol.* **51**, 1711–1738 (2010).
- Shaw, A. M., Hauri, E. H., Fischer, T. P., Hilton, D. R. & Kelley, K. A. Hydrogen isotopes in Mariana arc melt inclusions: implications for subduction dehydration and the deep-Earth water cycle. *Earth Planet. Sci. Lett.* **275**, 138–145 (2008).
- Müller, R. D., Sdrolias, M., Gaina, C. & Roest, W. R. Age, spreading rates, and spreading asymmetry of the world's ocean crust. *Geochem. Geophys. Geosyst.* **9**, Q04006 (2008).
- Emry, E. L., Wiens, D. A. & Garcia-Castellanos, D. Faulting within the Pacific plate at the Mariana trench: implications for plate interface coupling and subduction of hydrous minerals. *J. Geophys. Res.* **119**, 3076–3095 (2014).
- Oakley, A. J., Taylor, B. & Moore, G. F. Pacific plate subduction beneath the central Mariana and Izu-Bonin fore arcs: new insights from an old margin. *Geochem. Geophys. Geosyst.* **9**, Q06003 (2008).
- Christensen, N. I. Serpentinites, peridotites, and seismology. *Int. Geol. Rev.* **46**, 795–816 (2004).
- Nishimura, C. E. & Forsyth, D. W. The anisotropic structure of the upper mantle in the Pacific. *Geophys. J. Int.* **96**, 203–229 (1989).
- Feng, H. S.-H. *Seismic Constraints on the Processes and Consequences of Secondary Igneous Evolution of Pacific Oceanic Lithosphere*. PhD thesis, Massachusetts Institute of Technology and Woods Hole Oceanographic Institution (2016).
- Faccenda, M., Gerya, T. V., Mancktelow, N. S. & Moresi, L. Fluid flow during slab unbending and dehydration: implications for intermediate-depth seismicity, slab weakening and deep water recycling. *Geochem. Geophys. Geosyst.* **13**, Q01010 (2012).
- Reynard, B. Serpentine in active subduction zones. *Lithos* **178**, 171–185 (2013).
- Nakatani, T. & Nakamura, M. Experimental constraints on the serpentinization rate of fore-arc peridotites: implications for the upwelling condition of the slab-derived fluid. *Geochem. Geophys. Geosyst.* **17**, 3393–3419 (2016).
- Korenaga, J. On the extent of mantle hydration caused by plate bending. *Earth Planet. Sci. Lett.* **457**, 1–9 (2017).
- David, C., Wong, T.-F., Zhu, W. & Zhang, J. Laboratory measurement of compaction-induced permeability change in porous rocks: implications for the generation and maintenance of pore pressure excess in the crust. *Pure Appl. Geophys.* **143**, 425–456 (1994).
- Stein, C. A. & Stein, S. A model for the global variation in oceanic depth and heat flow with lithospheric age. *Nature* **359**, 123–129 (1992).
- Schwartz, S. et al. Pressure–temperature estimates of the lizardite/antigorite transition in high pressure serpentinites. *Lithos* **178**, 197–210 (2013).
- Ji, S. et al. Seismic velocities, anisotropy, and shear-wave splitting of antigorite serpentinites and tectonic implications for subduction zones. *J. Geophys. Res.* **118**, 1015–1037 (2013).
- Miller, N. C. & Lizarralde, D. Finite-frequency wave propagation through outer rise fault zones and seismic measurements of upper mantle hydration. *Geophys. Res. Lett.* **43**, 7982–7990 (2016).
- Emry, E. L. & Wiens, D. A. Incoming plate faulting in the northern and western Pacific and implications for subduction zone water budgets. *Earth Planet. Sci. Lett.* **414**, 176–186 (2015).
- Nakanishi, M., Tamaki, K. & Kobayashi, K. Magnetic anomaly lineations from Late Jurassic to Early Cretaceous in the west-central Pacific Ocean. *Geophys. J. Int.* **109**, 701–719 (1992).
- Takahashi, N., Kodaira, S., Tatsumi, Y., Kaneda, Y. & Suyehiro, K. Structure and growth of the Izu-Bonin-Mariana arc crust: 1. Seismic constraint on crust and mantle structure of the Mariana arc-back-arc system. *J. Geophys. Res.* **113**, B01104 (2008).
- Hayes, G. P., Wald, D. J. & Johnson, R. L. Slab1.0: a three-dimensional model of global subduction zone geometries. *J. Geophys. Res.* **117**, B01302 (2012).
- Emry, E. L., Wiens, D. A., Shiobara, H. & Sugioka, H. Seismogenic characteristics of the northern Mariana shallow thrust zone from local array data. *Geochem. Geophys. Geosyst.* **12**, Q12008 (2011).

Acknowledgements We thank P. J. Shore, H. Jian and the captains, crew and science parties of the RVs *R. Revelle* and *Melville* for data collection; S. Wei and M. Pratt for helping with data processing; R. Parai and M. J. Krawczynski for discussions; and X. Wang for support. IRIS PASSCAL and OBSIP provided land-based seismic instrumentation and ocean-bottom seismographs, respectively. This work was supported by the GeoPRISMS Program under NSF grant OCE-0841074 (D.A.W.).

Reviewer information *Nature* thanks C. Rodríguez Ranero & D. Shillington for their contribution to the peer review of this work.

Author contributions C.C. and M.E., advised by D.A.W., analysed the seismic data. W.S. developed and modified the Monte Carlo inversion code. C.C. and D.A.W. took the lead in writing the manuscript, and all authors discussed the results and edited the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0655-4>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0655-4>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to C.C.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data processing and group- and phase-velocity tomography. The data used in the group- and phase-velocity tomography were collected mainly by 19 ocean-bottom seismographs and seven temporary island-based seismic stations deployed from January 2012 to February 2013 (Fig. 1a). The distribution of ocean-bottom seismographs covers the outer-rise region of the trench and the Mariana forearc region. In addition, we used data from three island stations from the USGS Northern Mariana Islands Seismograph Network that were active over the same time period.

We carried out ambient noise tomography (ANT) following previously described procedures^{34,35}. The daily vertical-component seismograms were corrected for instrument responses and clock errors, and down-sampled to two samples per second. We applied running-average time-domain normalization and spectral whitening to minimize the effects of large earthquakes. Seismograms from all station pairs were then cross-correlated and stacked over the entire time period of the deployment. Frequency-time analysis^{34,36} was applied to the symmetric components of the stacked cross-correlations to measure Rayleigh-wave group and phase velocities between periods of 8 s and 25 s. For each frequency, only station pairs with distances larger than twice the wavelength were kept. All dispersion curves were screened to exclude those with inconsistent measurements at adjacent periods. For each period, a ray-theory-based tomography method³⁷ was applied to dispersion measurements with signal-to-noise ratios greater than 5 to produce Rayleigh group- and phase-velocity maps on a grid of nodes spaced at 0.2°. The tomographic inversion returns the isotropic and azimuthal anisotropic components of the Rayleigh-wave group and phase velocity (Extended Data Fig. 3).

We applied a Helmholtz tomography (HT) method³⁸ to teleseismic Rayleigh waveforms to determine phase velocities at longer periods. From the International Seismological Centre (ISC) catalogue, we selected seismograms from 380 earthquakes with surface-wave magnitudes (M_S) larger than 4.5 and epicentral distances between 25° and 150° that occurred during the time when the stations were operating (Extended Data Fig. 4). The raw seismogram of each event was cut from the time of origin of the earthquake to 12,000 s after it. Before any further analysis, the vertical-component seismograms were down-sampled to one sample per second and instrument responses were corrected. Noise in seismograms at long periods (>50 s) due to ocean swell and associated water-pressure variations, as well as tilt caused by local currents, were removed by correcting the vertical channel using horizontal and pressure channels^{39–41}.

This implementation of the HT method recovers frequency-dependent phase and amplitude information via the narrow-band filtering of the broadband cross-correlations between the vertical-component seismogram from a given station and the time-windowed seismograms from all other nearby stations. The phase delays and amplitude information were determined by fitting the narrow-band-filtered cross-correlations with a Gaussian wavelet³⁸. To eliminate the influence of poor-quality records, we estimated the coherence between waveforms from nearby stations for a series of periods from 21 s to 53 s, and only included those measurements with coherence larger than 0.5. For each earthquake and each period, we inverted the phase delays for spatial variations in dynamic phase velocity via the Eikonal equation⁴². We then further corrected the propagation effect via HT⁴³, producing maps of structure phase velocity with a spacing of 0.2°. This tomographic method returns the azimuthal isotropic phase velocity and the azimuthal anisotropic component at each node simultaneously.

Bayesian Monte Carlo inversion. We combined the ANT and HT results to provide more complete measurements of phase velocity for the vertically polarized S wave (SV wave) velocity inversion (Extended Data Fig. 5). The two sets of dispersion curves were combined in the geographic region that was well resolved by both methods. Phase velocities were interpolated onto a uniform grid of nodes with a spacing of 0.2° before being combined at each node. For phase velocities from ANT (8–25 s), the uncertainties were normalized at each period so that the uncertainty of the best-resolved node is 0.075 km s^{−1}. The uncertainties of the group velocity (8–21 s) were normalized so that the best-resolved node has an uncertainty of 0.188 km s^{−1}, 2.5 times the value for phase velocities⁴⁴. For phase velocities from HT (21–53 s), the uncertainties were normalized at each period so that the best-resolved node has an uncertainty equal to the standard deviation of velocity differences between HT results and results from a two-plane-wave tomography method⁴⁵ (Extended Data Fig. 6). We used a linear weighting average method to combine phase velocity measurements and uncertainty estimates for overlapping periods (22–25 s). A running average was then applied to make the resulting dispersion curve smoother. Group velocity results from ANT (8–21 s) were also included for the SV-wave velocity inversion to better fit water thickness and to improve resolution for shallower structure.

We use a Bayesian Monte Carlo algorithm⁴⁶ to invert the azimuthally averaged SV-wave velocity at each node. This approach allows us to apply prior constraints on crustal thickness and other parameters in a systematic way, to avoid

any potential bias of the starting model⁴⁷ and to derive formal estimates of velocity uncertainty.

The Bayesian Monte Carlo method constructs an a priori distribution of SV-wave velocity models at each node, defined by perturbations relative to the starting model and model constraints. Each model consists of four layers on top of a half-space: (1) water with starting thickness from bathymetry⁴⁸ that has been smoothed with a Gaussian filter (at a length of 125 km) and an allowed perturbation of ±1.5 km; (2) sediments; (3) crust; and (4) upper mantle from the Moho to a depth of 180 km. The sedimentary layer is described by two parameters: a layer thickness of 0.5 km with an allowed perturbation of ±0.5 km and a constant V_{SV} of 2.0 km s^{−1} with a perturbation of ±1.0 km s^{−1}. The crust is assumed to have linearly increasing velocity with depth and is described by three parameters: a layer thickness, and V_{SV} at the top and bottom of the layer. For the incoming plate east of the trench, the crustal thickness is allowed to vary by ±1.5 km around the starting value of 6.5 km. The forearc crustal thickness perturbs within 3 km, with starting values from a previous seismic refraction survey³¹ at the southern edge of the study region ranging from 19 km to 6.5 km. The top and bottom crustal V_{SV} are set at 3.0 km s^{−1} and 3.2 km s^{−1}, respectively, with a perturbation of ±1.0 km s^{−1}. The upper-mantle V_{SV} is parameterized by a *B*-spline, which is defined by seven nodes, with a perturbation of ±30% for the first five and of ±20% for the last two. We impose the constraint that the jumps in V_{SV} from the sediment to the crust and from the crust to the mantle are positive.

We also apply a physical dispersion correction with a reference period of 1 s (ref. ⁴⁹) using a one-dimensional attenuation (Q) model simplified from a seismic attenuation study in the same region⁵⁰. Compared to the Preliminary reference Earth model, our one-dimensional Q model for the forearc has a high-attenuation layer in the uppermost mantle: $Q_S = 60$ from the Moho to a depth of 100 km. For the incoming plate region, the uppermost mantle is set to have a typical lithospheric attenuation: $Q_S = 300$ from the Moho to a depth of 100 km.

For each grid node, the best-fitting model is identified and models are accepted if their χ^2 misfit is less than 50% higher than that of the best-fitting model^{44,46}. We also exclude models with mantle velocity higher than 4.9 km s^{−1}. The posterior distribution thus provides statistical information on all possible SV-wave velocity models that satisfy the Rayleigh-wave dispersion curves within tolerances depending on data uncertainties. An average model is then calculated from all accepted models and used for plotting and interpretations^{46,51}. Examples of the SV-wave velocity inversion at four representative nodes are shown in Extended Data Fig. 5. The results of the Bayesian Monte Carlo inversion fit the measured group- and phase-velocity dispersion curves well.

Robustness of the thick low-velocity layer. The application of the Bayesian Monte Carlo algorithm⁴⁶ helps to avoid the potential bias of the starting models and provides better prior constraints on crustal thickness and other parameters. However, it uses a *B*-spline method to parameterize the upper-mantle V_{SV} , and so may smooth a thin low-velocity layer over a wider depth range. Here we run simulations to test whether the thick low-velocity region observed above and below the surface of the subducting slab can be caused by smoothing of the 6-km-thick lower-velocity crust as a result of the inversion parameterization and the lack of precise depth resolution.

For the target node, we set up a one-dimensional shear-velocity model without a low-velocity serpentinized slab mantle layer, based on our prior knowledge of the geometry of the Mariana subduction zone (Extended Data Fig. 1a), and calculate the synthetic phase and group dispersion curves⁵². We then apply the Bayesian Monte Carlo inversion with the same parameterizations as in our study to the synthetic phase and group data and obtain a one-dimensional reconstructed shear-velocity structure. Examples for two nodes are shown in Extended Data Fig. 1b, c, and suggest that the thick low-velocity layer observed in our study cannot be caused purely by smearing of the subducting oceanic crust.

Serpentine and water-filled cracks. Previous studies at various convergent margins generally attribute the observed upper-mantle slow velocity anomalies to the presence of serpentine^{6,53–57}. Although the three main serpentine minerals—lizardite, chrysotile and antigorite—have the same water content (about 13 wt%), their physical properties, including seismic velocity^{17,27} and stability field^{58,59}, are different, owing to the different crystal structure. Lizardite and chrysotile are the more abundant serpentine minerals in hydrated mantle rocks formed at low temperatures and are stable up to about 320°C at 1 GPa. When the temperature reaches between 320°C and 390°C, lizardite is progressively replaced by antigorite at the grain boundaries and in the core of the lizardite meshes²⁶. Antigorite is the main stable serpentine mineral at higher temperature (up to about 620°C at 1 GPa)^{26,59–62}. Lizardite and chrysotile have much lower shear velocities (roughly 2.3 km s^{−1}) compared to antigorite (roughly 3.7 km s^{−1}) at 600 MPa (ref. ²⁷). At 600 MPa, the experimental relationship between shear-velocity V_S and serpentine volume fraction (Φ) is $V_S = 4.51 - 2.19\Phi$ for lizardite and chrysotile, and $V_S = 4.51 - 0.84\Phi$ for antigorite²⁷. For the same reduction in shear velocity, the serpentinization percentage (and thus the water content) estimated assuming

an antigorite component will be roughly 2.5 times that assuming a lizardite and chrysotile component²⁷. This feature makes it imperative to decide which serpentine minerals are present before making any further interpretations of velocity reductions. To estimate a lower bound for the serpentinization percentage, we use 4.51 km s^{-1} as the reference velocity for the unaltered mantle instead of the 4.7 km s^{-1} that we observe (Fig. 2a–c).

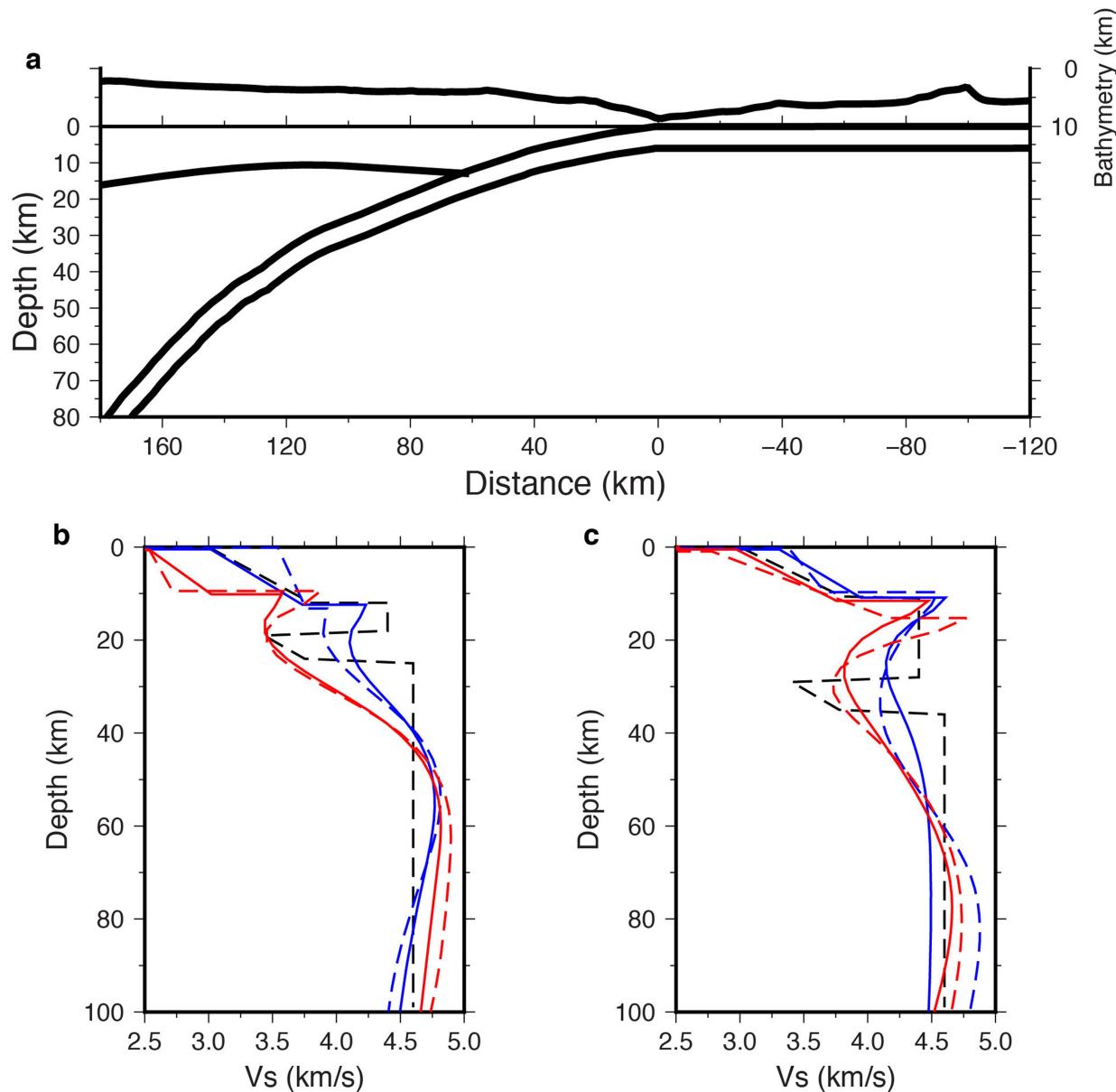
It has been argued²³ that the same reduction in velocity can also be caused by water-filled porosity, without involving substantial bulk hydration. This argument presumes a non-fractured isotropic media, which is incompatible with the field observations in the Mariana subduction zone, where numerous normal faults and normal-fault earthquakes have been detected^{15,16}. Instead, a porous media with aligned cracks is the more appropriate assumption^{63,64}. On the other hand, this argument can be applied to the slab only before subduction. When the slab starts to subduct, the confining pressure increases with increasing depth, causing closure of cracks and expulsion of free water within these cracks and/or porosity²⁴. Thus, this hypothesis is not applicable to the velocity reduction observed within the slab at greater depth after subduction.

Estimation of global subduction-zone water flux. We recalculated the global water flux into the subduction zone on the basis of a previous estimate³, by re-evaluating the water content in the slab mantle. Because serpentine minerals are stable up to 620°C , young and warm subducting plates have less potential to be serpentinized to great depth. According to thermal models for oceanic plates²⁵, only plates older than about 40 Myr have their 600°C isotherm deeper than about 30 km beneath the seafloor; we therefore set 40 Myr as an age threshold for the subducting plate to be affected by deeper serpentinization. For subduction zones with subducting plates younger than 40 Myr, we take the water-flux estimates from ref. ³. For subduction zones with incoming plates older than 40 Myr, we assume that the slab mantle is partially serpentinized (2 wt% water) to 20 km below the Moho and keep the water volume in the sediment and crust as in ref. ³. This rough estimate suggests that the global subduction-zone water flux should increase to $3.0 \times 10^9 \text{ Tg Myr}^{-1}$. **Anisotropy effect of serpentine along bending faults.** The anisotropic effects of serpentine distributed along bending faults could lead to an overestimate of the serpentinization percentage²⁸. We calculated the long-wavelength azimuthal anisotropy produced by evenly distributed serpentine layers⁶⁵. According to our estimate of the serpentinization percentage (roughly 19 vol%), we assume that pure serpentine layers (450 m thick) were evenly distributed within isotropic peridotite with a spacing of 2 km. We show results for two layering geometries: vertical layering and 45° dipping layering (Extended Data Fig. 2). Serpentine layers were set to have the following properties: isotropic compression-wave velocity $V_p = 5.1 \text{ km s}^{-1}$, isotropic shear-wave velocity $V_s = 2.32 \text{ km s}^{-1}$ and density $\rho = 2.52 \text{ g cm}^{-3}$. Peridotite layers were set to have the following properties: $V_p = 8.1 \text{ km s}^{-1}$, $V_s = 4.51 \text{ km s}^{-1}$ and $\rho = 3.32 \text{ g cm}^{-3}$. The azimuthally averaged quasi-SV-wave velocity for the case of 45° dipping layering is 4.08 km s^{-1} , very close to the Voigt average SV-wave velocity (4.1 km s^{-1}) that we use to estimate the serpentinization percentage directly. This result suggests that the anisotropy effect of serpentine distributed along bending faults may be less important when estimating serpentinization percentage²⁸, especially when the bending faults are dipping.

Data availability

Raw seismic data are available at the Data Management Center of the Incorporated Research Institutions for Seismology (<http://www.iris.edu/dms/nodes/dmc>) under network IDs MI and XF. Network and station information can be found at the IRIS website (<http://www.ds.iris.edu/mda>).

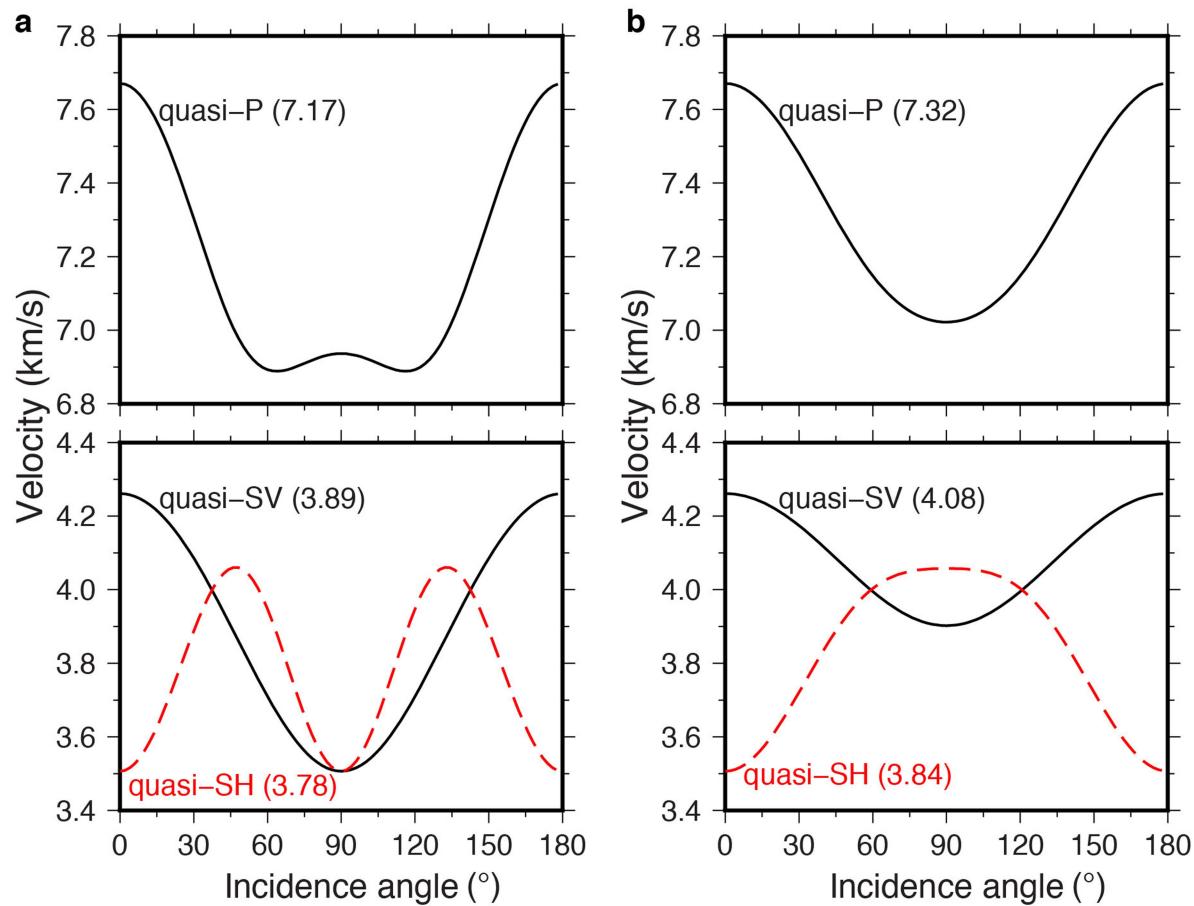
34. Bensen, G. D. et al. Processing seismic ambient noise data to obtain reliable broad-band surface wave dispersion measurements. *Geophys. J. Int.* **169**, 1239–1260 (2007).
35. Lin, F.-C., Moschetti, M. P. & Ritzwoller, M. H. Surface wave tomography of the western United States from ambient seismic noise: Rayleigh and Love wave phase velocity maps. *Geophys. J. Int.* **173**, 281–298 (2008).
36. Levshin, A. L. & Ritzwoller, M. H. Automated detection, extraction, and measurement of regional surface waves. *Pure Appl. Geophys.* **158**, 1531–1545 (2001).
37. Barmin, M. P., Ritzwoller, M. H. & Levshin, A. L. A fast and reliable method for surface wave tomography. *Pure Appl. Geophys.* **158**, 1351–1375 (2001).
38. Jin, G. & Gaherty, J. B. Surface wave phase-velocity tomography based on multichannel cross-correlation. *Geophys. J. Int.* **201**, 1383–1398 (2015).
39. Bell, S. W., Forsyth, D. W. & Ruan, Y. Removing noise from the vertical component records of ocean-bottom seismometers: results from year one of the Cascadia Initiative. *Bull. Seismol. Soc. Am.* **105**, 300–313 (2015).
40. Crawford, W. C. & Webb, S. Identifying and removing tilt noise from low-frequency ($<0.1 \text{ Hz}$) seafloor vertical seismic data. *Bull. Seismol. Soc. Am.* **90**, 952–963 (2000).
41. Webb, S. C. & Crawford, W. C. Long-period seafloor seismology and deformation under ocean waves. *Bull. Seismol. Soc. Am.* **89**, 1535–1542 (1999).
42. Lin, F.-C., Ritzwoller, M. H. & Snieder, R. Eikonal tomography: surface wave tomography by phase front tracking across a regional broad-band seismic array. *Geophys. J. Int.* **177**, 1091–1110 (2009).
43. Lin, F.-C. & Ritzwoller, M. H. Helmholtz surface wave tomography for isotropic and azimuthally anisotropic structure. *Geophys. J. Int.* **186**, 1104–1120 (2011).
44. Shen, W. et al. A seismic reference model for the crust and uppermost mantle beneath China from surface wave dispersion. *Geophys. J. Int.* **206**, 954–979 (2016).
45. Yang, Y. & Forsyth, D. W. Regional tomographic inversion of the amplitude and phase of Rayleigh waves with 2-D sensitivity kernels. *Geophys. J. Int.* **166**, 1148–1160 (2006).
46. Shen, W., Ritzwoller, M. H., Schulte-Pelkum, V. & Lin, F.-C. Joint inversion of surface wave dispersion and receiver functions: a Bayesian Monte-Carlo approach. *Geophys. J. Int.* **192**, 807–836 (2013).
47. Wei, S. S. et al. Seismic evidence of effects of water on melt transport in the Lau back-arc mantle. *Nature* **518**, 395–398 (2015).
48. Lindquist, K. G., Engle, K., Stahlke, D. & Price, E. Global topography and bathymetry grid improves research efforts. *Eos* **85**, 186 (2004).
49. Kanamori, H. & Anderson, D. L. Importance of physical dispersion in surface wave and free oscillation problems: review. *Rev. Geophys.* **15**, 105–112 (1977).
50. Pozgay, S. H., Wiens, D. A., Conder, J. A., Shiobara, H. & Sugioka, H. Seismic attenuation tomography of the Mariana subduction system: implications for thermal structure, volatile distribution, and slow spreading dynamics. *Geochem. Geophys. Geosyst.* **10**, Q04X05 (2009).
51. Wei, S. S. et al. Upper mantle structure of the Tonga-Lau-Fiji region from Rayleigh wave tomography. *Geochem. Geophys. Geosyst.* **17**, 4705–4724 (2016).
52. Herrmann, R. B. Computer programs in seismology: an evolving tool for instruction and research. *Seismol. Res. Lett.* **84**, 1081–1088 (2013).
53. Contreras-Reyes, E., Grevermeyer, I., Flueh, E. R., Scherwath, M. & Heesemann, M. Alteration of the subducting oceanic lithosphere at the southern central Chile trench-outer rise. *Geochem. Geophys. Geosyst.* **8**, Q07003 (2007).
54. Contreras-Reyes, E. et al. Deep seismic structure of the Tonga subduction zone: implications for mantle hydration, tectonic erosion, and arc magmatism. *J. Geophys. Res.* **116**, B10103 (2011).
55. DeShon, H. R. & Schwartz, S. Y. Evidence for serpentinization of the forearc mantle wedge along the Nicoya Peninsula, Costa Rica. *Geophys. Res. Lett.* **31**, L21611 (2004).
56. Garth, T. & Rietbrock, A. Constraining the hydration of the subducting Nazca plate beneath Northern Chile using subduction zone guided waves. *Earth Planet. Sci. Lett.* **474**, 237–247 (2017).
57. Savage, B. Seismic constraints on the water flux delivered to the deep Earth by subduction. *Geology* **40**, 235–238 (2012).
58. Evans, B. W. The serpentine multisystem revisited: chrysotile is metastable. *Int. Geol. Rev.* **46**, 479–506 (2004).
59. Guillot, S., Schwartz, S., Reynard, B., Agard, P. & Prigent, C. Tectonic significance of serpentinites. *Tectonophysics* **646**, 1–19 (2015).
60. Perrillat, J.-P. et al. Kinetics of antigorite dehydration: a real-time X-ray diffraction study. *Earth Planet. Sci. Lett.* **236**, 899–913 (2005).
61. Ulmer, P. & Trommsdorff, V. Serpentine stability to mantle depths and subduction-related magmatism. *Science* **268**, 858–861 (1995).
62. Wunder, B. & Schreyer, W. Antigorite: high-pressure stability in the system $\text{MgO}-\text{SiO}_2-\text{H}_2\text{O}$ (MSH). *Lithos* **41**, 213–227 (1997).
63. Gurevich, B. Elastic properties of saturated porous rocks with aligned fractures. *J. Appl. Geophys.* **54**, 203–218 (2003).
64. Hudson, J. A., Liu, E. & Crampin, S. The mechanical properties of materials with interconnected cracks and pores. *Geophys. J. Int.* **124**, 105–112 (1996).
65. Backus, G. E. Long-wave elastic anisotropy produced by horizontal layering. *J. Geophys. Res.* **67**, 4427–4440 (1962).



Extended Data Fig. 1 | Robustness test of the low-velocity zone.

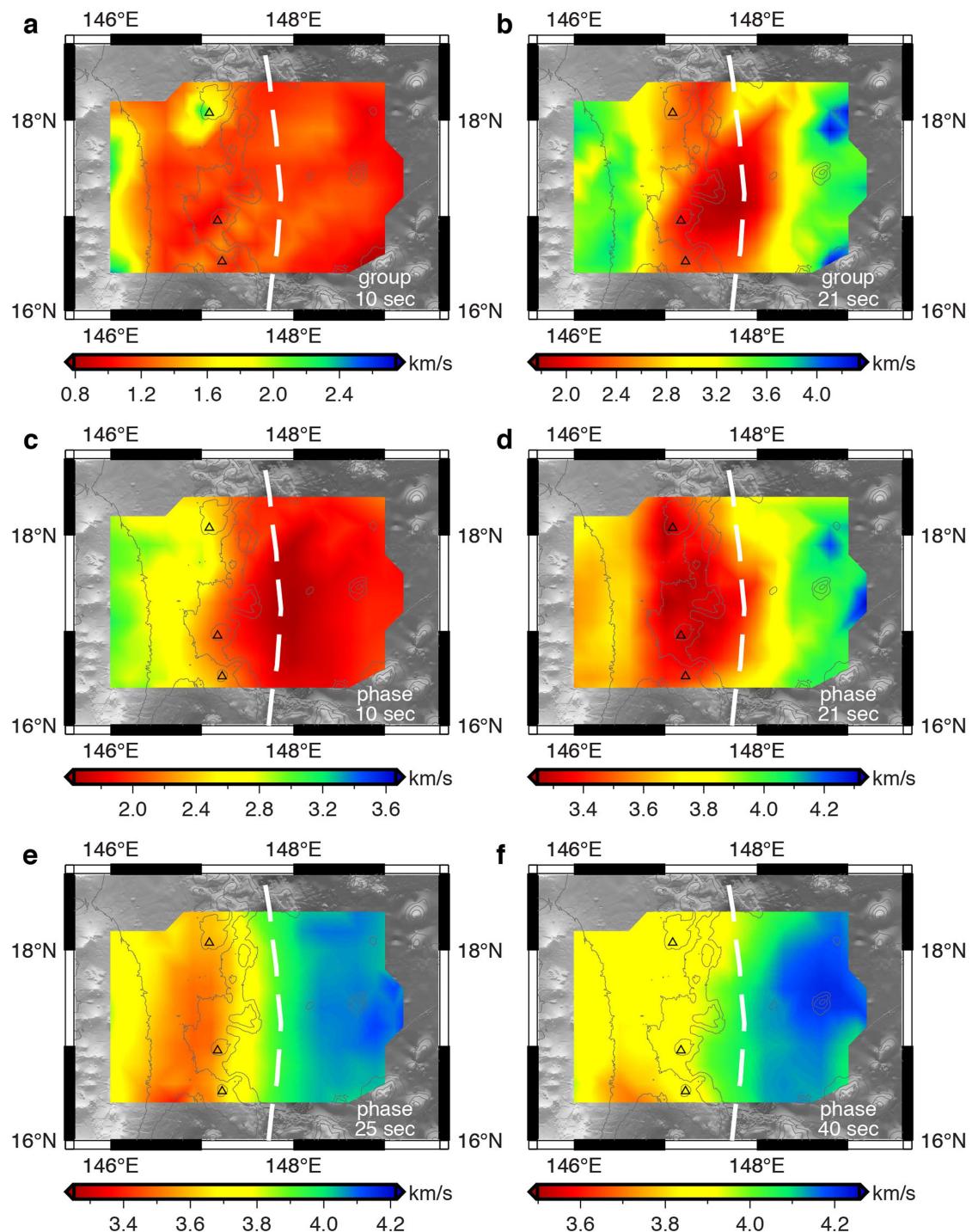
a, The assumed geometry of the subduction zone according to our prior knowledge. **b, c**, Simulation results for nodes 80 km (**b**) and 110 km (**c**) landward from the trench. The black dashed lines are the input one-

dimensional models; blue dashed and solid lines are the best-fitting and average models from the Monte Carlo inversion of the synthetic dispersion curves, respectively; red dashed and solid lines are the best-fitting and average models from the Monte Carlo inversion of the real data.



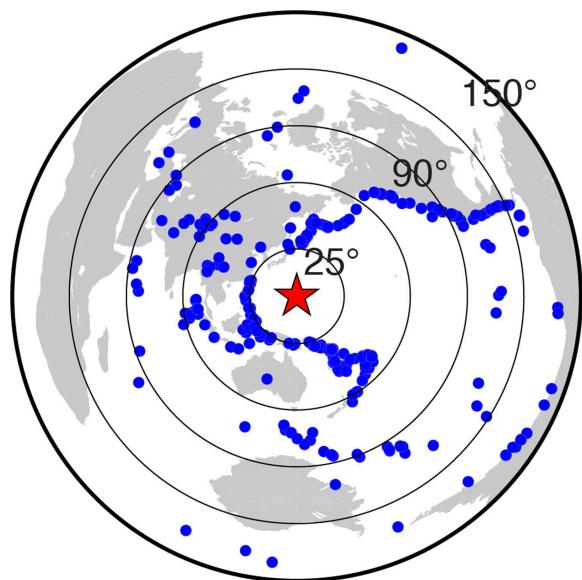
Extended Data Fig. 2 | Azimuthal anisotropy from evenly distributed serpentine layers (of thickness 450 m and with a spacing of 2 km).
a, Result for vertical layering. **b**, Result for 45° dipping layering. Numbers

in the parenthesis are the mean velocity for quasi-P, quasi-SV or quasi-SH. The incidence angle is defined relative to the strike of the layer: 0° is parallel and 90° is normal to the strike.

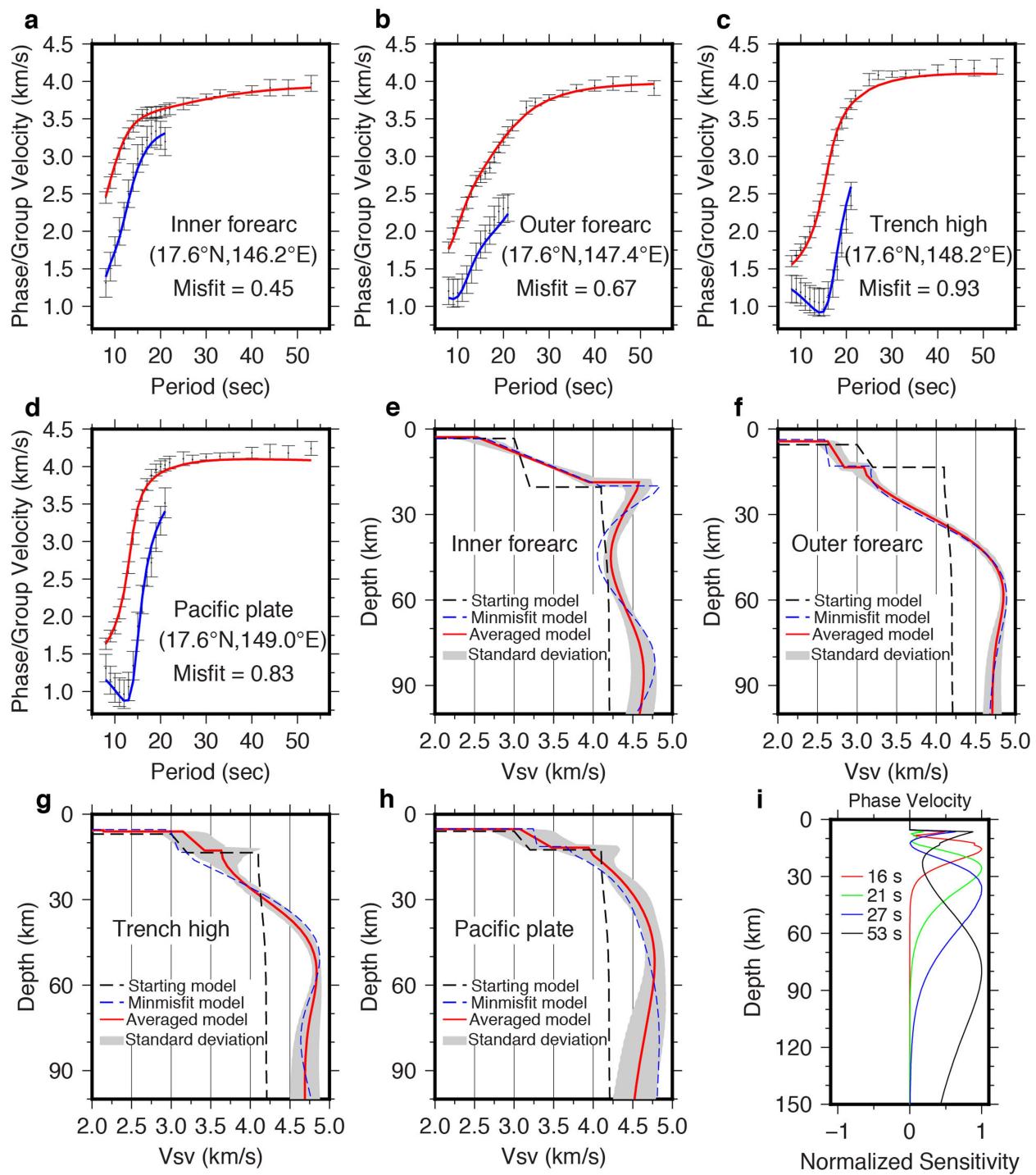


Extended Data Fig. 3 | Maps of azimuthally averaged group and phase velocity. **a, b**, Group velocity (colour scale) at periods of 10 s (a) and 21 s (b) inverted by ANT. **c, d**, Phase velocity (colour scale) at periods of 10 s (c) and 21 s (d) from ANT. **e, f**, Phase velocity (colour scale) for periods

of 25 s (e) and 40 s (f) inverted by HT. 3-km, 4-km and 5-km bathymetry contours are shown as thin grey lines. The trench axis and serpentine seamounts are shown as in Fig. 1a.

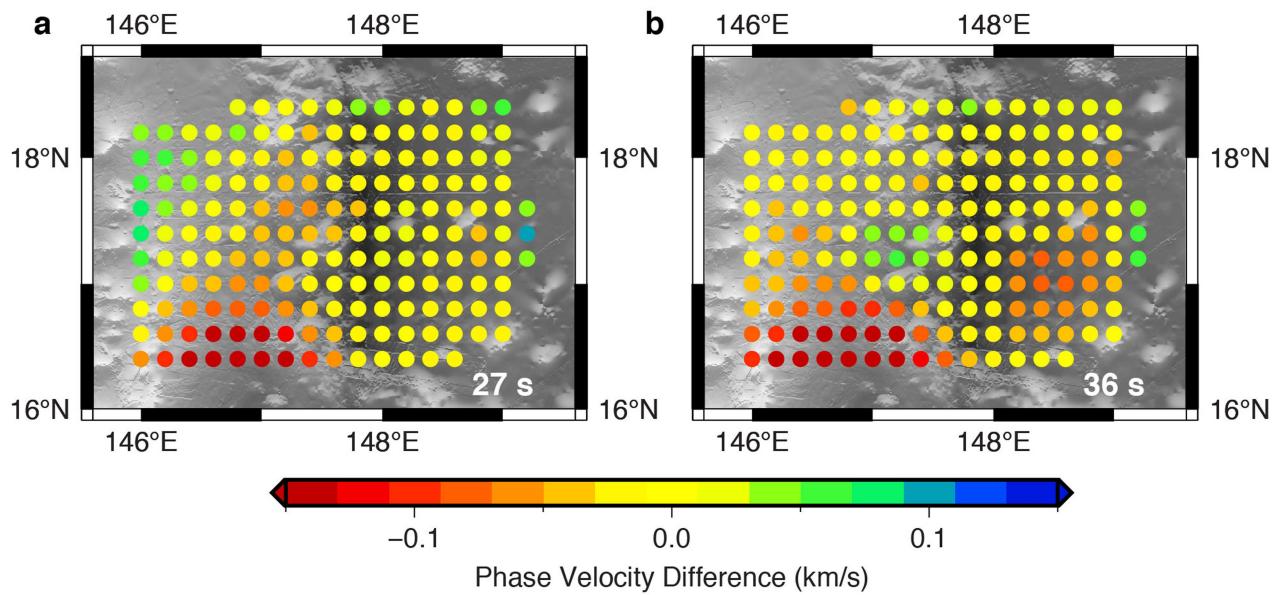


Extended Data Fig. 4 | Earthquakes used in this study. Blue dots represent ISC earthquake locations. The red star shows the location of the Mariana trench.



Extended Data Fig. 5 | Examples of Monte Carlo inversion and phase-velocity sensitivity kernel. **a-d**, The joint Rayleigh phase and group dispersion data (error bars, one standard deviation) and computed phase (red solid lines) and group (blue solid lines) dispersion curves from the Bayesian Monte Carlo averaged model, for four locations: **a**, inner forearc;

b, outer forearc; **c**, trench high; **d**, Pacific plate. **e-h**, Shear-velocity model from the Bayesian Monte Carlo inversion for the four example locations. **i**, Phase-velocity sensitivity kernels at example periods, calculated using the average velocity model in **g**.



Extended Data Fig. 6 | Comparison between Rayleigh-wave isotropic phase velocities determined from teleseismic tomography using HT and a two-plane-wave method. a, At 27 s. b, At 36 s.

Remarkable muscles, remarkable locomotion in desert-dwelling wildebeest

Nancy A. Curtin^{1*}, Hattie L. A. Bartlam-Brooks¹, Tatjana Y. Hubel¹, John C. Lowe¹, Anthony R. Gardner-Medwin², Emily Bennett³, Stephen J. Amos¹, Maja Lorenc¹, Timothy G. West¹ & Alan M. Wilson^{1*}

Large mammals that live in arid and/or desert environments can cope with seasonal and local variations in rainfall, food and climate¹ by moving long distances, often without reliable water or food en route. The capacity of an animal for this long-distance travel is substantially dependent on the rate of energy utilization and thus heat production during locomotion—the cost of transport^{2–4}. The terrestrial cost of transport is much higher than for flying (7.5 times) and swimming (20 times)⁴. Terrestrial migrants are usually large^{1–3} with anatomical specializations for economical locomotion^{5–9}, because the cost of transport reduces with increasing size and limb length^{5–7}. Here we used GPS-tracking collars¹⁰ with movement and environmental sensors to show that blue wildebeest (*Connochaetes taurinus*, 220 kg) that live in a hot arid environment in Northern Botswana walked up to 80 km over five days without drinking. They predominantly travelled during the day and locomotion appeared to be unaffected by temperature and humidity, although some behavioural thermoregulation was apparent. We measured power and efficiency of work production (mechanical work and heat production) during cyclic contractions of intact muscle biopsies from the forelimb flexor carpi ulnaris of wildebeest and domestic cows (*Bos taurus*, 760 kg), a comparable but relatively sedentary ruminant. The energetic costs of isometric contraction (activation and force generation) in wildebeest and cows were similar to published values for smaller mammals. Wildebeest muscle was substantially more efficient (62.6%) than the same muscle from much larger cows (41.8%) and comparable measurements that were obtained from smaller mammals (mouse (34%)¹¹ and rabbit (27%)). We used the direct energetic measurements on intact muscle fibres to model the contribution of high working efficiency of wildebeest muscle to minimizing thermoregulatory challenges during their long migrations under hot arid conditions.

We tested the hypothesis that wildebeest undertake long-range locomotion from grazing sites to water sources and that their muscles are optimized to deliver a low cost of transport (COT). We chose blue wildebeest living in the Makgadikgadi Pans National Park in Botswana, because water is sparse and is found in known locations and grazing is limited.

Wildebeest were captured by darting from a helicopter and fitted with tracking collars of our own design¹⁰, which contained GPS, a 3D accelerometer, 3D gyroscope, 3D magnetometer, a humidity sensor and a black globe thermometer¹² (to measure combined effect of solar radiation, air temperature and air velocity on the animal) (Fig. 1a and Methods). Collar mass was 1,050 g, which is 0.5% of body mass. After 18 months, collars released automatically, drop-off failures were recovered by re-darting, and 17 of the 20 deployed collars were recovered (to date) (Extended Data Table 1). During tranquilization, a muscle biopsy with aponeurosis at each end was removed from the flexor carpi ulnaris (a forelimb flexor) muscle of six wildebeest (leg length to serratus ventralis insertion 1.09 m, approximate body mass, 220 kg) under open aseptic conditions and immediately returned to the field

laboratory by helicopter. Equivalent biopsies were collected from flexor carpi ulnaris of seven adult cows (leg length to serratus insertion 1.28 m, 770 kg) at a UK abattoir. Work was approved by the Ethics & Welfare Committee of the Royal Veterinary College (RVC 2013 1233).

In the dry season, ranging wildebeest drank (exclusively) from the Boteti river, usually every 2–3 days; however, thirteen out of sixteen individuals went four days between drinking events at least once, and seven individuals went five days (Fig. 1b, c, g, i, j, n). The wildebeest grazed 5–15 km away from the river and covered 20–40 km between drinks (Fig. 1d, f, j) and they consistently drank in the middle of the day (time clusters around multiples of 24 h, Fig. 1g, j, m). Each year, wildebeest migrated to and from the wet season range, 60–80 km over three or four days across a waterless environment. This daily distance (Fig. 1e, f, h, i) was further than most non-migrating wildebeest travelled.

Humidity peaked in the wet season and dropped below 12% in the dry season (median for September, Fig. 1o). Globe temperature on the collar peaked in October (Fig. 1p). Neither humidity nor globe temperature appeared to curtail the maximum distance travelled (Fig. 1k, l). Data from two fixed weather stations on the walking route (Fig. 1b) recorded considerably ($5.6 \pm 0.6^\circ\text{C}$ (mean \pm s.d.)) higher peak and lower minimum ($3.6 \pm 1.1^\circ\text{C}$) globe temperatures than the animal collars (Extended Data Fig. 1), indicating that animals showed behavioural thermoregulation at both temperature extremes. The mean daily maximum, by month, exceeded 38°C in nine out of twelve months.

These data show that wildebeest are able to cover up to 80 km and last up to five days without drinking, contrary to published observations that report a requirement for daily drinking¹³. This pattern of activity is necessary in the studied environment, because of the depletion of grazing areas close to the only permanent water source as well as the biannual migration across a waterless environment between seasonal ranges. This behaviour persisted even in the late dry season, when daytime temperatures were high and low humidity would maximize respiratory water loss.

Almost all locomotion (97% of total distance) was at a slow walk within a narrow speed range centred around a preferred speed of 1.14 m s^{-1} and a stride frequency centred around 1.00 Hz (Fig. 1q, r). This corresponds to a dimensionless speed of 0.35 (leg length 1.09 m), which is around the optimum for mechanical COT minimization^{14,15}. Occasional short bursts of faster locomotion at up to 15 m s^{-1} occurred with approximately double the stride frequency (peak of 1.92 Hz at 4 m s^{-1}) accounting for 3% of the daily distance travelled. The intermediate gait of trotting was very rarely used: only 5% of the faster non-walking strides (0.15% of total).

We cycled muscle biopsies at 0.5 Hz at 25°C , which approximated the temperature-corrected walking stride frequency of wildebeest (Fig. 1r) and recorded force, length and heat production. We systematically varied the stimulation duration (duty cycle) and stimulation phase (start of stimulation relative to start of shortening; Extended Data Fig. 2). Figure 2 shows the stimulus pattern that elicited high power

¹Structure & Motion Laboratory, Royal Veterinary College, University of London, Hatfield, UK. ²Department of Neuroscience, Physiology & Pharmacology, University College London, London, UK.

³Okavango Research Institute, University of Botswana, Maun, Botswana. *e-mail: ncurtin@rvc.ac.uk; awilson@rvc.ac.uk

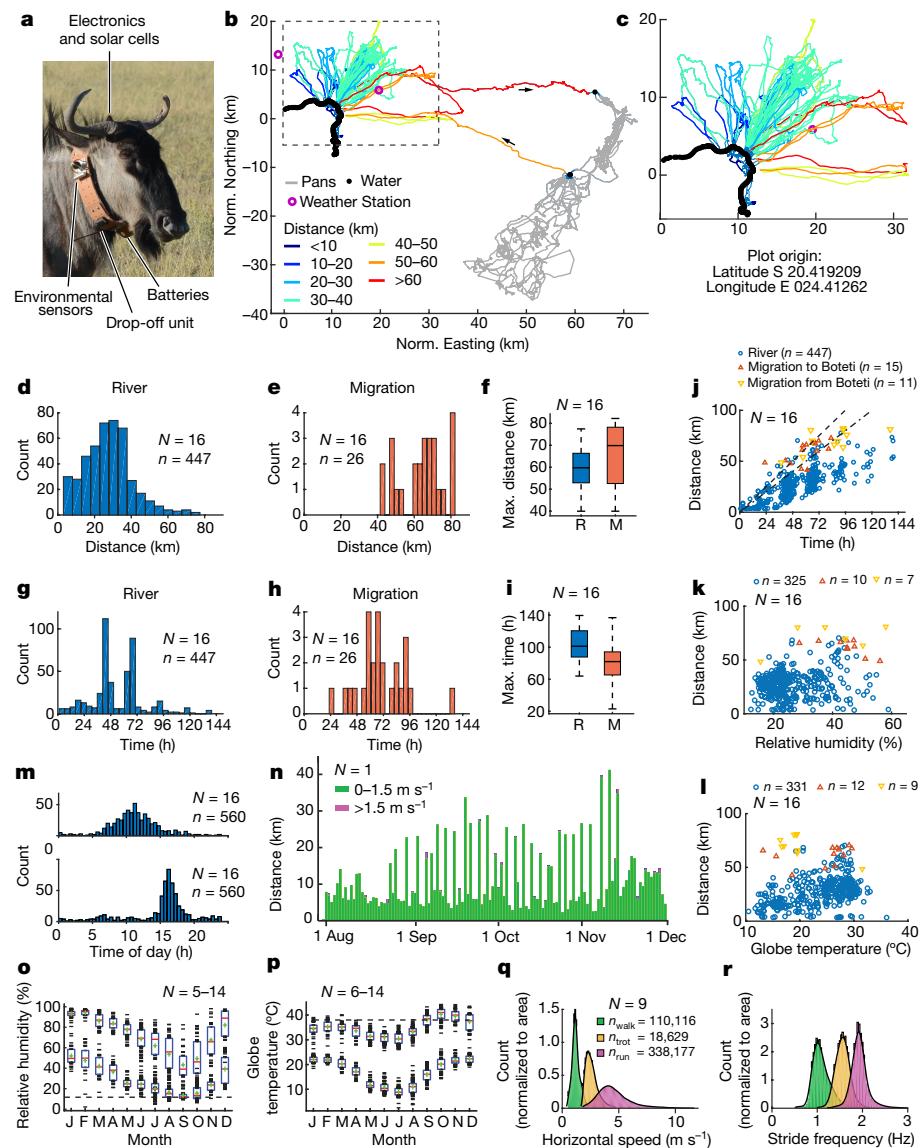


Fig. 1 | Locomotion of desert wildebeest. **a**, A collared wildebeest. **b**, The typical range of a wildebeest. The black line indicates the Boteti river grazing ground forays are coloured by distance between drinking events. Migratory journeys between dry and wet season range are colour-coded by distance between drinks. Journeys within the wet season range are indicated in grey. The data are normalized (Norm.) so that zero is on the plot not on the equator intersection with Greenwich meridian. **c**, Dry season range. **d, e**, Distance covered between drinks at the river (**d**) and during migrations (**e**). N , number of animals, n = number of journeys. **f**, Distribution of longest distances between drinks for each individual. Data are median, interquartile range and range. **g, h**, Time between drinks showing a circadian drinking pattern. **i**, Data as in **f** for the longest individual times between drinks. **j**, Distance covered versus time taken for each journey, dashed lines represent 20 and 25 km per day. **k**, Distance

and efficiency from a wildebeest fibre bundle. Successful experiments were performed on five fibre bundles from four wildebeest and five bundles from five cows.

Power (Fig. 3a, e) increased with stimulation duration (between a duty cycle of 0.1 and 0.3) and the power curve shifted to the left and became narrower. Peak power was produced when stimulation and shortening started at the same time (phase zero) (n numbers are shown in Extended Data Table 2; individual data points are shown in Extended Data Fig. 3).

Impulse (an integral of time and active stress) varied with duty cycle and with phase (Fig. 3c, g) and the largest impulse was produced under isometric conditions. When more of the stimulation occurred during

between drinks against mean humidity during the period. Blue, dry range; red, migration to the river; yellow, migration from the river. **l**, Data as in **k** for mean temperature. **m**, Time of arrival at, and departure from, the river. **n**, Daily distance for one wildebeest during dry season showing a pattern of a long walk to/from the river interspersed with one or two grazing days. Green, walking; purple, faster locomotion. **o**, Daily maximum and minimum ambient humidity of the median of all animal-mounted sensors, grouped by month. Red line, median, the '+' symbol indicates the mean, boxes are the interquartile range and individual values are shown as horizontal dashes. **p**, Daily maximum and minimum globe temperature derived as in **o**. **q, r**, Speeds (**q**) and stride frequencies (**r**) that were used derived from high-rate data only (nine collars, number of strides are shown) and subdivided into gaits with normalization for equal area under curves for each gait. Green, walk; orange, trot; purple, canter/gallop.

stretch (and phase became more negative), the impulse was higher (Fig. 3c, g) and the cost per unit of impulse lower (Fig. 3d, h). For wildebeest, the minimum cost per unit of impulse during cyclic movement ($0.056 \pm 0.011 \text{ J kg}^{-1} \text{ kPa}^{-1} \text{ s}^{-1}$ (mean \pm s.e.m.), $n = 4$) was slightly less than for isometric contraction ($0.068 \pm 0.021 \text{ J kg}^{-1} \text{ kPa}^{-1} \text{ s}^{-1}$, $n = 4$), although the stimulation duration was different (Extended Data Table 3a). The cost of impulse under isometric conditions for wildebeest muscle was similar to that derived from published data from rats and our data from cows (Extended Data Table 3b).

Three fibre bundles from wildebeest were used to perform a fatiguing series of contractions (see Methods), which reduced force considerably (Fig. 3i), but force-producing capacity was completely restored after

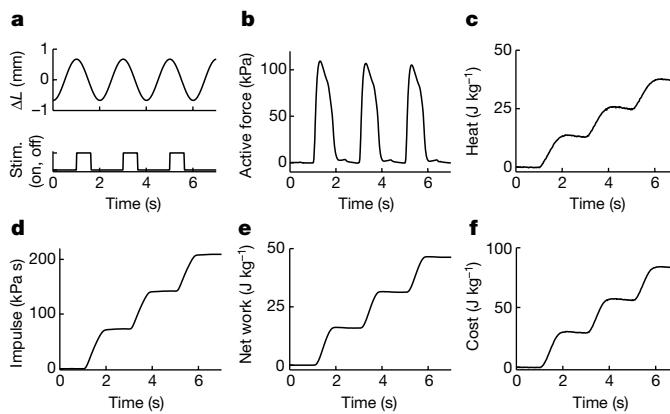


Fig. 2 | Example records. Data for wildebeest are from bundle 4A, duty cycle 0.3, phase 0. **a**, Length change (ΔL) and imposed stimulus (Stim.) pattern. **b, c**, Active stress (**b**) and heat produced by the fibre bundle (**c**). **d–f**, Quantities calculated from **a–c**. **d**, Impulse is the integral of active stress and time. **e**, Net work is the integral of active force and length change. **f**, Cost was evaluated as the total energy = heat production + net work. For three cycles, the heat was 36.22 (**c**), net work was 46.27 (**e**) and heat + net work was 82.50 (**f**). Therefore efficiency = net work/(heat + net work) = 56.1% (46.27/82.50). This efficiency point is shown in Extended Data Fig. 4d (a square at phase 0). Active force is expressed relative to the cross-sectional area of the bundle, and heat, work and cost are expressed relative to the mass of the bundle. Muscle fibre length was 7.1 mm, fibre bundle mass was 9.02 mg. Net work over three cycles was 46.3 J kg^{-1} , an average cycle power of 7.7 W kg^{-1} .

a recovery period without stimulation (Fig. 3j). This resilience of the muscles is high compared with equivalent data for other species and it may reflect the low metabolic demands of the fibres and/or good intra-muscle buffering of pH. The muscles were observed to be very red in appearance, suggesting substantial oxidative capacity as observed histochemically in the black wildebeest (*Connochaetes gnou*)¹⁶.

For both wildebeest and cows, the efficiency of the production of net work (defined as positive work minus any negative work) was strongly dependent on stimulation phase and peaked at either -0.1 or 0 for all duty cycles (Fig. 3b, f). Efficiency values are enthalpy efficiencies—that is, the mechanical power divided by the sum of heat production rate and mechanical power (Extended Data Fig. 4 and Extended Data Table 4). Furthermore, they are initial enthalpy efficiency values,

meaning that the enthalpy is produced during the contractions and is from ATP and phosphocreatine use, coupled by the creatine kinase reaction. The mean maximum efficiencies that we measured for wildebeest muscle ($62.6 \pm 2.3\%$ (mean \pm s.e.m., $n = 4$)) and for cow muscle ($41.8\% \pm 1.2$ ($n = 5$)) are high. The efficiency of the wildebeest muscle is higher than the locomotor muscle of any currently measured species except for the muscle of tortoises¹⁷ (see Extended Data Table 5), an animal fabled for slow locomotor speed and good endurance. Tortoises have a long stance times and a low stride frequency¹⁸. The cow is by far the largest mammal—and the wildebeest the only long-distance locomotion specialist—that have been the subject of such energetic measurements. On the basis of their larger body mass, cows would be predicted to have a COT that is 44% lower than wildebeest⁵.

The mechanisms that are responsible for the high initial enthalpy efficiency could arise from (1) the ATP-driven Ca^{2+} uptake into the sarcoplasmic reticulum and (2) the cross-bridge cycle in which one ATP is used in each round of attachment, work-producing filament sliding and detachment. We used the principles set out in a previous study¹⁹ (see Methods) to compare the work that the cross-bridge actually does with the amount of work it could theoretically do; this is the efficiency of the fundamental contractile event. We explored whether, mechanistically, the cross-bridges attach and detach at appropriate locations to yield maximum work. In tortoise muscle, each cross-bridge cycle produces 45.8 zJ of work ($z = 10^{-21}$), amounting to 92% of the theoretically possible 50 zJ , while the remaining 8% was released as heat¹⁷ (see Extended Data Table 6) and the wildebeest cross-bridge generates work that amounts to between 74 and 86% of the theoretically possible 50 zJ , with the remainder dissipating as heat. For cows, the corresponding calculation gives cross-bridge work between 50 and 57% of the theoretical maximum. The high efficiency of the cross-bridge cycle in wildebeest muscle indicates that work is produced with relatively low heat production. Therefore, the efficiency of the cross-bridge cycle itself has an important role, supporting the ability of the wildebeest to function effectively in its environment. The use of a slow walking gait with high duty cycle and low stride frequency probably exploits this efficiency to the full.

The energy that is required (above resting metabolism) for a 220-kg wildebeest walking 20 km per day (Fig. 1j) would be around 0.379 MJ km^{-1} or 7.58 MJ per day, requiring an additional oxygen consumption⁵ of 18.91 km^{-1} (Methods). Assuming a 5% oxygen extraction from air²⁰, this requires an additional ventilation of 378 l km^{-1} and a consequent additional water loss in expired air of 15.2 ml km^{-1} or

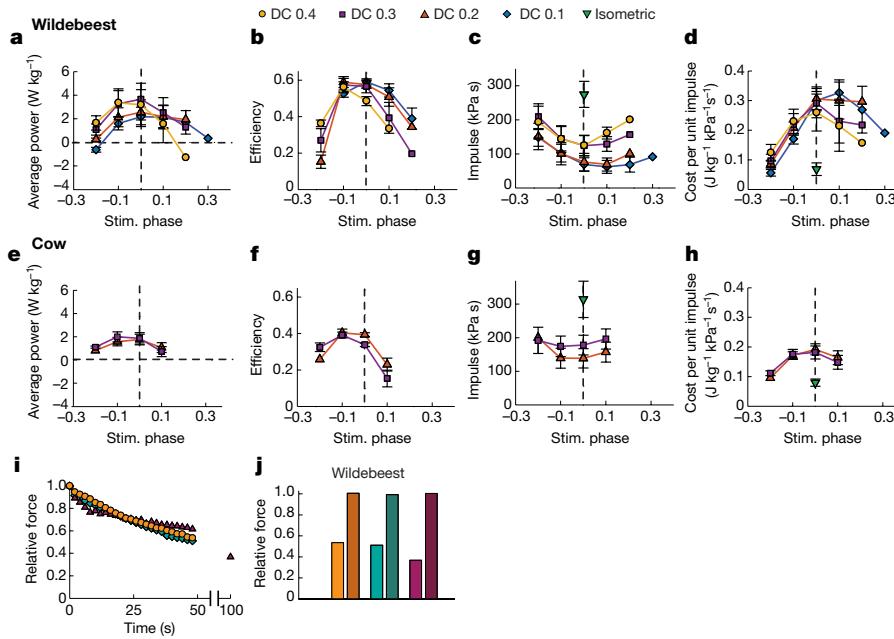


Fig. 3 | Muscle mechanical and energetic performance. **a–d**, Data from wildebeest. **e–h**, Data from cows. Relationship between stimulus phase and mechanical and energetic outputs during three cycles of movement at 0.5 Hz and different stimulus duty cycles (DC). Isometric contractions at DC = 0.4 are shown as a green downward triangle. **a, e**, Average power. **b, f**, Efficiency is the power per rate of heat + work output. **c, g**, Impulse is the integral of active stress and time. **d, h**, Cost per unit of impulse is the (heat + work) per impulse. Data are mean \pm s.e.m., $n = 3–5$. Data are mean only, when $n = 2$. n numbers are shown in Extended Data Table 2. **i**, Fatigue of fibre bundles from wildebeest. Peak force in 25 or 50 isometric contractions (at 2-s intervals, DC = 0.4). Peak force is normalized to the first contraction. **j**, Extent of fatigue and recovery. Peak force of the last contraction (light colours) and after the recovery period of 10 and 30 min (dark colours). The three bundles are shown by different colours.

304 ml per day. This is mostly offset by 11.5 ml km⁻¹ (230 ml per day) water that is generated through metabolism, leading to a net water loss of only 3.7 ml km⁻¹ or 74 ml per day.

Heat accumulation can be a major issue for a hot desert-dwelling animal^{1,21}. Without heat dissipation, the heat energy from walking 20 km (7.58 MJ) would raise body temperature by an intolerable 10°C. When ambient temperature exceeds body temperature (Fig. 1p and Extended Data Fig. 1), thermoregulation can only be achieved by evaporative cooling and, for horizontal locomotion, almost all energy used by muscles converts—directly or indirectly—to heat within the body of the animal.

The behavioural thermoregulation reported above suggests that wildebeest avoided direct heat—for example, by standing in the shade or in moving air—but this may not be concomitant with long-distance movement. An additional way to minimize temperature problems would be to move at dawn, dusk or overnight, yet nearly all long-distance movements occurred during warm daylight hours, arriving at the river late morning and leaving mid-afternoon (Fig. 1m). This behaviour possibly reduces the risk of predation by lions²².

To dissipate all the extra heat (7.58 MJ) generated by walking 20 km by evaporation, the 220-kg wildebeest would need to evaporate 3.36 l of water per day. This is substantially more than the 0.23 l per day generated chemically from the extra metabolism. The net loss (3.13 l per day; 1.4% of body mass per day) would be a considerable dehydration load. If the same work in transport were performed by muscle with cow muscle enthalpy efficiency (41.8% versus 62.6%), the energetic cost of work, respiratory and thermoregulatory water loss would all be 50% greater (4.7 l per day; 2.1% of body mass per day). Additional water depletion will occur associated with basal metabolism, possibly as much as 5.4% of body mass per day—the observed water intake of grazing wildebeest²³. Animals become debilitated through dehydration when they reach around 20% body mass loss²⁴, corresponding to between three and four days between drinks. We suggest that wildebeest, particularly during migration, may gain a substantial increase in range under hot arid conditions as a result of having highly efficient muscles.

In summary, wildebeest, although not considered extreme arid environment specialists, can undertake long journeys in the absence of water in hot dry conditions and frequently spend three and occasionally up to five days without drinking. This requires them to have a low COT, which is likely to be delivered, in part, by muscles that are specialized at the level of the cross-bridge to deliver more mechanical work and release less heat from each ATP molecule split than any mammalian muscle studied to date. Equivalent data for cow muscle shows that the muscle specialization is not simply attributable to animal size. The economical muscles are therefore likely to be critical in minimizing heat accumulation and enabling the nomadic lifestyle of these wildebeest and exploitation of distant pastures, particularly during periods of nutritional deprivation and/or high ambient temperature. Such physiological adaptations may be critical to surviving in a changing world.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0602-4>.

Received: 12 April 2018; Accepted: 24 August 2018;

Published online 24 October 2018.

1. Schmidt-Nielsen, K. *Desert Animals: Physiological Problems of Heat and Water* (Clarendon, Oxford, 1965).
2. Hedenstrom, A. Optimal migration strategies in animals that run: a range equation and its consequences. *Anim. Behav.* **66**, 631–636 (2003).
3. Hein, A. M., Hou, C. & Gillooly, J. F. Energetic and biomechanical constraints on animal migration distance. *Ecol. Lett.* **15**, 104–110 (2012).
4. Tucker, V. A. The energetic cost of moving about: walking and running are extremely inefficient forms of locomotion. Much greater efficiency is achieved by birds, fish—and bicyclists. *Am. Sci.* **63**, 413–419 (1975).

5. Taylor, C. R., Heglund, N. C. & Maloiy, G. M. O. Energetics and mechanics of terrestrial locomotion. I. Metabolic energy consumption as a function of speed and body size in birds and mammals. *J. Exp. Biol.* **97**, 1–21 (1982).
6. Kram, R. & Taylor, C. R. Energetics of running: a new perspective. *Nature* **346**, 265–267 (1990).
7. Wilson, A. M., Watson, J. C. & Lichtwark, G. A. Biomechanics: a catapult action for rapid limb protraction. *Nature* **421**, 35–36 (2003).
8. Alexander, R., Maloiy, G. M. O., Njau, R. & Jayes, A. Mechanics of running of the ostrich (*Struthio camelus*). *J. Zool.* **187**, 169–178 (1979).
9. Alexander, R., Maloiy, G. M. O., Ker, R., Jayes, A. & Warui, C. The role of tendon elasticity in the locomotion of the camel (*Camelus dromedarius*). *J. Zool.* **198**, 293–313 (1982).
10. Wilson, A. M. et al. Locomotion dynamics of hunting in wild cheetahs. *Nature* **498**, 185–189 (2013).
11. Barclay, C. J. Efficiency of fast- and slow-twitch muscles of the mouse performing cyclic contractions. *J. Exp. Biol.* **193**, 65–78 (1994).
12. Hetem, R. S., Maloney, S. K., Fuller, A., Meyer, L.C. & Mitchell, D. Validation of a biotelemetric technique, using ambulatory miniature black globe thermometers, to quantify thermoregulatory behaviour in ungulates. *J. Exp. Zool.* **307A**, 342–356 (2007).
13. Estes, R. *The Behavior Guide to African Mammals* Vol. 64 (Univ. California Press, Berkeley, 1991).
14. Kuo, A. D. A simple model of bipedal walking predicts the preferred speed-step length relationship. *J. Biomech. Eng.* **123**, 264–269 (2001).
15. Bertram, J. E. Constrained optimization in human walking: cost minimization and gait plasticity. *J. Exp. Biol.* **208**, 979–991 (2005).
16. Kohn, T. A., Curry, J. W. & Noakes, T. D. Black wildebeest skeletal muscle exhibits high oxidative capacity and a high proportion of type IIx fibres. *J. Exp. Biol.* **214**, 4041–4047 (2011).
17. Woledge, R. C. The energetics of tortoise muscle. *J. Physiol.* **197**, 685–707 (1968).
18. Williams, T. L. Experimental analysis of the gait and frequency of locomotion in the tortoise, with a simple mathematical description. *J. Physiol.* **310**, 307–320 (1981).
19. Barclay, C. J. Energetics of contraction. *Compr. Physiol.* **5**, 961–995 (2015).
20. Butler, P. J. et al. Respiratory and cardiovascular adjustments during exercise of increasing intensity and during recovery in thoroughbred racehorses. *J. Exp. Biol.* **179**, 159–180 (1993).
21. Hetem, R. S., Maloney, S. K., Fuller, A. & Mitchell, D. Heterothermy in large mammals: inevitable or implemented? *Biol. Rev. Camb. Philos. Soc.* **91**, 187–205 (2016).
22. Valeix, M. et al. Behavioral adjustments of African herbivores to predation risk by lions: spatiotemporal variations influence habitat use. *Ecology* **90**, 23–30 (2009).
23. MacFarlane, W. V., Howard, B., Haines, H., Kennedy, P. & Sharpe, C. M. Hierarchy of water and energy turnover of desert mammals. *Nature* **234**, 483–484 (1971).
24. Maloiy, G. M. O. Water economy of the Somali donkey. *Am. J. Physiol.* **219**, 1522–1527 (1970).

Acknowledgements We thank R. Woledge for contributing to early design of experiments; C. Barclay for helping us to fabricate the thermocouple elements; our field assistants, N. Terry and M. Claase; A. R. Wilson for logistical support and editorial contributions; M. Flyman (Department of Wildlife and National Parks) for his support and enthusiasm and J. O'Connor and P. O'Riordan (Dawn Meats, Bedford) for enabling cow muscle collection. Funding was provided by the EPSRC (EP/H013016/1), BBSRC (BB/J018007/1) and ERC (323041). A Botswana Research Permit EWT 8/36/4 was held by A.M.W. and A.M.W. was a registered Botswana veterinarian.

Reviewer information *Nature* thanks J. E. A. Bertram, R. Hetem and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions A.M.W., N.A.C. and H.L.A.B.-B. conceived, designed and led the study; H.L.A.B.-B. and E.B. led and organized field work. A.M.W. performed veterinary procedures and biopsies. N.A.C., A.R.G.-M., M.L. and T.G.W. undertook muscle experiments and N.A.C. and A.R.G.-M. analysed and interpreted muscle data. J.C.L., A.M.W. and S.J.A. designed and built collars and weather stations. T.Y.H., A.M.W. and H.L.A.B.-B. analysed and interpreted collar data, A.M.W. made the water balance model and A.M.W. and N.A.C. wrote paper with input from all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0602-4>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0602-4>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to N.A.C. or A.M.W.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Animals. Custom-made GPS collars were deployed on twenty female wildebeest that were collared in their wet season range in the Makgadikgadi Pan National Park in Botswana in May 2016. Animals were free-darted from a Robinson R44 helicopter by A.M.W., a veterinary surgeon registered in Botswana, using 5 mg etorphine hydrochloride (M99, Novartis), 80 mg azaperone (Kyron Laboratories), 1,700 IU lyophilized hyalase (Kyron Laboratories). For reversal, diprenorphine (Novartis) and/or naltrexone (Kyron Laboratories) were used. Timed-release drop-offs (TRD, Lotek) released collars automatically after 18 months. Two collars remain deployed at time of writing, 16 collars were retrieved and two lost. Data were recorded for 17 animals for variable durations, because of difficulties downloading large datasets, collar failure and animal death (Extended Data Table 1). Of all drop-offs, 59% failed and those collars were recovered by re-darting from a helicopter.

Open biopsies were obtained under aseptic conditions from the flexor carpi ulnaris muscle. Animals were treated with a long acting antibiotic (Norocillin LA, Norbrook, 6 mg kg⁻¹ benzathine penicillin and 4.5 mg kg⁻¹ procaine penicillin, intramuscular injection) and a nonsteroidal anti-inflammatory drug (Metacam, Boehringer Ingelheim, 0.5 mg meloxicam per kg body weight, intramuscular injection). The flexor carpi ulnaris muscle was chosen because it is superficial, accessible, unipennate and a few millimetres thick. The fibres run between aponeurosis sheets and the fibre length was appropriate for mounting in the thermopile.

Collars. Collars used in this study were designed, engineered and assembled at the RVC^{10,25} (Fig. 1a). Collars were fitted with a VHF tracking transmitter (African Wildlife Tracking), one lithium polymer rechargeable battery (Active Robots), charged by a solar cell array that consists ten monocrystalline silicon solar cells (KXOB22-12X1, Ixys) and two standby D cell lithium thionyl chloride batteries (LSH20T, Saft Groupe SA). They were further equipped with a GPS module (M8N GPS module; u-Blox AG), a six-axis inertial measurement unit incorporating a three-axis accelerometer and a three-axis gyroscope (MPU-6050, TDK-Invensense), a separate low power three-axis accelerometer (MMA8652, NXP Semiconductors), a three-axis magnetometer (HMC5883, Honeywell International) and an ambient light sensor (TSL2591, Ams AG). Additionally, the collars were outfitted with an environmental measurement unit that recorded humidity, ambient temperature and globe temperature every 30 s. The unit consisted of a humidity sensor (HIH6131, Honeywell) and two temperature sensors (MCP9808, Microchip Technology), one positioned near the humidity sensor inside a silvered, cross-drilled metal tube for ambient temperature measurement and the other mounted in the centre of a miniature blackened 30-mm globe (Press Spinning & Stamping) for derivation of mean radiant temperature¹². The unit was connected on an I2C bus via a gateway and protected power switch, which prevented any damage interfering with main collar operation.

GPS positions were recorded at 5-min intervals. During June, October and January (cold dry, hot dry and hot wet season, respectively), collars switched dynamically to a sample rate of 5 Hz (50 Hz inertial measurement unit (IMU)) during higher speed locomotion when triggered by the activity of the animal^{10,25}. The GPS module provides an estimation of each position and an instantaneous velocity data point (position accuracy: 0.95 m (median s.d.), horizontal speed accuracy: 0.30 m s⁻¹ (median s.d.)). During higher speed locomotion, IMU and GPS are Kalman fused to give substantially higher accuracy data¹⁰ (see 'Runs').

Fixed-position weather stations consisted of electronic boards and batteries (same as in the collars) mounted in clear polycarbonate cases with a temperature and a humidity sensor contained within a Stevenson Screen 3D printed of poly-lactic acid. These were mounted approximately 1.25 m off the ground (to match collar height) on dead trees at the locations shown on Fig. 1b. Data were recorded every 30 s.

Movement and location. Collar analysis was conducted in MATLAB (Mathworks). GPS position data with a horizontal accuracy estimate greater than 15 m were excluded from the analysis. Location was divided into an area close to the Boteti river, where most animals spend the dry season, with the river being the only water source, and the pans to the east, where they spend the wet season. Animals migrate from the river to the pans at the beginning of the wet season (November) and back to the Boteti river at the beginning of the dry season when the pans dry out (April–May). The direct distance between river and the closest edge of the pans is about 40–55 km. During the wet season, water is readily available in the pans, therefore we concentrated our analyses on times at the river and during migration between river and pans to calculate distance and time between drinking. The time that the animals spent in the pans were excluded from the analysis. One animal spent a whole year at the pans bringing the number of animals used for all analysis, except gait data, to 16.

The location of the river was derived from Google Earth by generating a path along the river and saving the coordinates as a CSV file.

Drinking was assumed to have occurred whenever the wildebeest was less than 500 m from the river. Distance and time between drinking was calculated based on the cumulative distance/time between two drinking events. Paths between drinking events were colour-coded based on the covered distance (Fig. 1 b, c) by summing displacement between 5-min GPS fixes, which will be an underestimate of the actual distance covered²⁶.

Migration events were identified manually and only analysed if the first or last drinking event in the pans could reliably be identified as a known water hole. Out of 47 migrations, these criteria were met and distance and time calculated for 15 migrations to the pans and 11 from the pans (Fig. 1e, h).

Maximum distance and maximum time between drinking were calculated for each individual during times at the river and migration; median values over all individuals are displayed in Fig. 1f, i.

Humidity and ambient temperature were interpolated to give a value that coincides with the 5-min GPS position and then averaged over the time between drinking events. Environmental sensors operated independently from the rest of the collar and there are times when no environmental data are available, reducing the number of river events and migration events from the original 447 and 26 to 325 and 17 (humidity) and 331 and 21 (temperature) (the *n* values for each are given Fig. 1k, l).

Collar environmental sensors shut down and restarted unexpectedly, possibly owing to moisture penetration into the electronics. However, a median humidity and ambient temperature over time was calculated from the collar data that were available at that time (data from between 1 and 14 individuals, typically 6–8). Minimum and maximum values were extracted using a peak detection algorithm and median daily maximum and minimum values were calculated for each month (Fig. 1o, p). Those medians over the course of the month contain data from 5–14 individuals for humidity and 6–14 individuals for ambient temperature.

Runs. Stride parameters, such as speed and frequency, were calculated from the triggered high-resolution data¹⁰. GPS-INS (Global Positioning System-Inertial Navigation System) processing was used to reduce noise and improve precision for the position and velocity analysis, as well as increase the temporal resolution of the data. GPS and IMU measurements were fused using a 12-state extended Kalman filter in loosely coupled architecture^{10,25}. For the combined GPS-IMU data, the position accuracy was estimated to be 0.30 m and speed accuracy to be 0.17 m s⁻¹.

Vertical accelerations were used to determine stride times^{10,25}. Stride frequency was calculated from the time between acceleration peaks and divided by two for symmetrical gaits. Horizontal stride speed was derived from the Kalman-filtered velocity averaged over each stride to remove the effects of speed fluctuation through the stride and collar oscillation relative to the centre of mass. Stride speed was weighted with the preceding and following stride to remove outliers^{10,27,28}. Gait was assigned based on speed thresholds: walking for speeds up to 1.8 m s⁻¹, trot between 1.8 m s⁻¹ and 4.7 m s⁻¹ and running above 4.7 m s⁻¹. Dimensionless speed¹⁴ was $\sqrt{(\text{leg length} \times 9.81)}$.

Muscle analyses of fibre bundles from wildebeest and cow. Measurement of energy use at the level of a muscle, rather than the level of the whole animal (oxygen consumption), is challenging. The most direct approach is to measure the mechanical work and heat released by a working fibre bundle during contraction. The individual temperature changes are small (around 0.001 °C) and rapid, and thus require highly sensitive custom-made thermopiles and a stable 'baseline' temperature. To characterize a muscle, an extensive series of measurements on living tissue with a viable cell membrane are required. Fibre bundles are dissected by hand and are particularly difficult to secure from large mammals, because most of their fibres are very long and always small in diameter (less than about 100 µm). Most published measurements on mammals have therefore been on laboratory rodents.

Fibre bundles from the flexor carpi ulnaris with aponeurosis (tendon) at each end were dissected from biopsies in saline (composition (mmol L⁻¹): NaCl 135, KCl 4.0, CaCl₂ 2.35, MgCl₂ 0.85, NaH₂PO₄ 1.0, NaHCO₃ 20 and glucose 5.5 and equilibrated with 95% O₂ and 5% CO₂). Aluminium foil T-shaped clips were attached to the tendons. The fibre preparation was mounted on a vertical thermopile between a fixed hook and a stainless-steel wire connected to the lever arm of a combined motor and force transducer (Series 300B Lever System and Series 400A Force Transducer System, Cambridge Technology).

Experiments were performed at 25 °C rather than at body temperature (38 °C) for consistency with other studies and because muscle preparations tend to be more resilient at lower temperatures. Muscle shortening velocity and power are temperature-dependent with a Q_{10} (the ratiometric increase in rate with a temperature increase of 10 °C) in the order of 2.3²⁹; this equates to a threefold increase from our measurements at 25 °C to the physiological temperature of 38 °C. The effect on optimum cycle frequency has not been defined, so we applied a more cautious twofold difference. Efficiency is temperature independent³⁰.

The fibre bundle was stimulated electrically (Isolated Stimulator Model DS2, Digitimer). Supra-maximal stimulus strength (at 60 Hz, 2 ms per pulse) and L_o , the fibre length giving maximum active force, were found for each fibre bundle at the

start of the experiment. The motor either held the fibre bundle at constant length (isometric) or imposed cycles of sinusoidal movement at 0.5 Hz, and peak-to-peak amplitude of about 18% L_o (Extended Data Fig. 2a). Stimulation consisted of three tetani, each lasting for part of the 2.0-s movement cycle (stimulation duty cycle (DC) = 0.1, and so on). The stimulation phase (timing of the tetanus within the movement cycle) was varied in steps of 0.1 of the movement cycle. See Extended Data Fig. 2 for values of movement amplitude, stimulation duty cycle and phase for wildebeest and cows. Passive force was recorded during sinusoidal movement without stimulation and was subtracted from the force produced during stimulation to give the active force. Isometric contractions were performed at intervals during the experiments. The duration of stimulation in these isometric contractions was 0.8 s, corresponding to duty cycle of 0.4 of the movement cycle duration for the fibre bundle.

Energetic measurements of wildebeest muscle. Muscle biopsies were obtained as described in the main text. Muscle temperature was measured by a custom-made thermopile (D1) consisting of antimony–bismuth thermocouples (Seebeck coefficient, 90.2 $\mu\text{V K}^{-1}$ per couple). The outputs from each of three 2-mm sections of the thermopile (8 couples per section) were recorded separately. A LabView program (National Instruments) controlled the stimulator and motor, and also recorded force, lever position and the outputs from the three thermopile sections. The program interfaced to the instruments using a USB-6229 DAQ (National Instruments).

Successful energetic measurements were made on five muscle fibre bundles from four wildebeest (Fig. 3, Extended Data Fig. 4 and Extended Data Table 4). Pooling the data for the two bundles from the same wildebeest yields a mean peak efficiency of 62.6% ($n=4$, range 57.3–66.6%). Removing the data for the bundle 3B (Extended Data Fig. 4f and Extended Data Table 4b), which had very low values for all parameters for unknown reasons for all parameters, gives a higher mean peak efficiency of 64.2% ($n=4$, range 60.2–66.6%) and averaging the second highest efficiency value determined for each of these four bundles give a mean of 62.6% (range 59.8–65.8) giving confidence that 62.6% is a robust measure of wildebeest muscle efficiency. Note that the peak efficiency of a fibre bundle is its highest value among all tested duty cycles and phases, whereas Fig. 3 shows the means for each combination of duty cycle and phase.

Energetic measurements of cow muscle. We used a thermopile, D2, which was similar to D1 described above. The Seebeck coefficient for D2 was 97.5 $\mu\text{V K}^{-1}$ per couple and D2 was longer and therefore more suitable for the longer fibre bundles from cows. Recordings were made from 1, 2 or 3 thermopile sections (8 couples per section, 2 mm per section), depending on the bundle length and on its position on the thermopile. Force was recorded as described for wildebeest.

Samples of the flexor carpi ulnaris were also retrieved from seven adult cows killed at a local abattoir. We report results for muscle samples from five of these cows (mean \pm s.d. estimated body mass 760 ± 23 kg, $n=5$; mean \pm s.d. leg length to top of scapula 1.43 ± 0.11 m, $n=5$). Leg length to insertion of serratus ventralis (same as wildebeest) was taken as 150 mm less, that is, 1.28 m. Muscle samples from two of the cows did not complete enough of the protocol to be included.

Fibre bundle size. After all of the recordings had been made, a digital photograph was made of the fibre bundle at L_o on the thermopile. The clip-to-clip length was measured from the image. The fibre bundle was pinned in a dissecting dish at the L_o clip-to-clip length and fixed in ethanol. The clips and tendon were removed and the fibre length was measured under the stereomicroscope. The fibre bundle was weighed after drying in room air. Dry mass/blotted mass was assumed to be same as we have measured for bundles of fibres from wild rabbit (0.188 ± 0.009 (mean \pm s.e.m.), $n=8$). Cross-sectional area was evaluated as blotted mass/ L_o .

Work, heat, efficiency and impulse. Recording for wildebeest and cow fibre bundles were analysed in the same way. Work was calculated as the integral of active force and length change. Impulse was calculated as the integral of active stress and time. Heat loss was evaluated using the time constant for heat loss measured using the Peltier method^{31,32}. Heat production was calculated from the thermopile output plus heat lost, using the Seebeck coefficient, number of thermocouples, and the heat capacity of the fibre bundle evaluated from its mass and a specific heat capacity of $3.668 \mu\text{J }^{\circ}\text{mK}^{-1} \text{mg}^{-1}$ muscle³³. Stimulus heat was measured after the muscle fibres had been made unexcitable with procaine (30 mmol l⁻¹ in saline). The heat values are reported as net of stimulus heat.

Work, heat and impulse produced in three complete cycles of movement (6 s) were measured. Work values are the net work, which is the summation of work done by the muscle during shortening minus the work done on the muscle by the motor during stretch. Power values are the net work in three cycles of movement/duration of three cycles; in other words, the average power produced during this 6-s period. Work, power and heat are reported per kg of muscle. Efficiency was evaluated as net work/heat + net work. Comparable efficiency results from the literature are shown in Extended Data Table 5. Impulse is reported as stress \times time (mN s mm⁻² = kPa s). The cost per unit impulse was compared with that of rat

muscle by performing equivalent calculations on previously published data³⁴ (see Extended Data Table 3b).

Comparable data were drawn from literature^{11,17,30,35–37} and are shown in Extended Data Table 5.

Fatigue and recovery. Tests of fatigue and recovery were done on some of the fibre bundles of wildebeest muscle. In our standard isometric protocol described above the fibre bundle performed three cycles consisting of 0.8 s stimulation followed by 1.2 s without stimulation. To examine fatigue and recovery of isometric stress we repeated 25 (or 50) cycles with stimulation, followed by a 10 (or 30) min recovery, then a test cycle with stimulation. Two sets of results from different fibre bundles are reported for the 25 stimulation cycles + 10-min recovery protocol, and one set of results for the 50 stimulation cycles + 30-min recovery protocol. After 25 contraction cycles, peak force fatigued to about 50% and after 50 cycles to 37% of its initial value (Fig. 3i). After the recovery period the initial force was completely restored; recovered/initial peak force = 1.001 ± 0.004 (mean \pm s.e.m.), $n=3$ (Fig. 3j).

Calculation of cross-bridge work from efficiency. The costs associated with Ca^{2+} reactions in wildebeest muscle may be low and contribute to high efficiency, but this could not be evaluated with the design of experiment that we used. To get some insight into the role of the cross-bridge cycle in efficiency, we have applied the principles set out in a previous study¹⁹ (Extended Data Table 6). Each cross-bridge cycle uses one ATP making 100 zJ ($z=10^{-21}$) of free energy available for conversion to mechanical work. However, measurements of the distance over which a cross-bridge is attached and the force it produces show that only 50 zJ of mechanical work can be done in a cross-bridge cycle. Such measurements have been made on muscle fibres from frog, dogfish and rat; the values are consistent and appear to be a fundamental property of the cross-bridges in vertebrate muscle. What varies between different muscles and muscles from different species is the fraction of the theoretically possible 50 zJ that is actually converted to work (Extended Data Table 6c). In mechanistic terms, a work per cross-bridge value less than 50 zJ means that when the cross-bridge attaches and detaches, it does so at locations that do not yield maximum work. In other words, the attached cross-bridge traverses only part of its force–length relationship in each cycle. We have calculated the cross-bridge work for wildebeest and cow muscle using our measured efficiency values and have assumed that the other required parameters are within the ranges measured for other muscles (Extended Data Table 6d, e).

Calculation of net COT, water and heat balance of a 220-kg wildebeest. COT data⁵ have been generated previously for a blue wildebeest with a body mass of 92 kg and an eland with a body mass of 213 kg and a regression line was derived from COT data for 11 species of artiodactyl. We used the regression line for the artiodactyls to get a whole animal net (that is, above resting metabolism) COT of 379 J m^{-1} for our 220-kg wildebeest. For comparison, the eland equation gave a net 220 kg whole-animal COT of 367 J m^{-1} and the slightly smaller 92 kg wildebeest value of 407 J m^{-1} , which were considered to be close enough to rely on the prediction using the regression line.

Taking an oxygen equivalent of $20.1 \text{ J per ml O}_2$ (20.1 kJ l^{-1})² gives an oxygen consumption of $0.379 \times 10^3/20.1 = 18.9 \text{ l O}_2$ per km. Assuming a 5% oxygen extraction from air (data for horses walking on a treadmill²⁰), this would require an additional ventilation of $18.9/0.05 = 378 \text{ l} = 0.378 \text{ m}^3$ per km. In this study system, locomotion occurred during hot, dry times of the day, so there is water loss owing to saturation of the extra ventilation with water in the lungs. Saturating that volume of air with water vapour at 38°C (body temperature), initial humidity of 12% (October median minimum humidity at about the same temperature, Fig. 1o) can be calculated as follows.

The maximum water content of air increases with temperature and can be sourced from engineering tables (for example, *Handbook of Chemistry and Physics*³⁸). At 38°C , it is $45.7 \text{ g H}_2\text{O m}^{-3}$ of air and taking 0.378 m^3 of air from 12% to 100% humidity requires $0.88 \times 45.7 \times 0.378 = 15.2 \text{ ml}$ of water (per km).

Respiratory water loss is substantially offset by metabolic production due to oxidation of carbohydrates ($0.0317 \text{ g H}_2\text{O per kJ}$) and fats ($0.0290 \text{ g H}_2\text{O per kJ}$). Using the mean of these (0.0304 g per kJ) yields 11.5 ml per km for water from 379 kJ per km for energy expenditure ($0.0304 \times 379 \times 1,000 = 11.5 \text{ ml H}_2\text{O per km walked}$).

Adding this to the respiratory water loss of 15.2 ml km^{-1} gives a net extra water loss due to locomotion of $-15.2 + 11.5 = -3.7 \text{ ml km}^{-1}$. Therefore, an additional loss of 74 ml day^{-1} (0.074 l day^{-1}) due to locomotion occurred when wildebeest walked 20 km per day.

Ambient temperature often exceeded 38°C between September and December (Fig. 1p and Extended Data Fig. 1) and the substantial radiant heat load in these environments will add to the requirement for evaporative cooling. When walking on level ground, most energy (heat + work) generated within muscles turns to heat within the animal—except for work done disturbing the environment. Dehydrated animals may store some heat during the day with a rise in temperature^{21,39}, allowing this stored heat to dissipate at night. Only small fluctuations of about 1°C are observed in wildebeest²¹. A 220-kg animal with a thermal capacity around 80% that

of water¹ ($3.34 \text{ kJ kg}^{-1} \text{ }^{\circ}\text{C}^{-1}$) could store only 0.73 MJ of heat with a $1\text{-}^{\circ}\text{C}$ rise, less than 10% of the 7.58 MJ generated by walking 20 km. Thus, daytime dissipation is essential, requiring evaporation of much more water than is evaporated with minimal ventilation.

Walking 20 km generates 7.58 MJ of heat. This heat would be dissipated by evaporating water (heat of vaporization $2,257 \text{ J g}^{-1}$). The volume to evaporate is equal to $7.58 \times 10^6 / 2,257 = 3,360 \text{ ml} = 3.36 \text{ l}$ of water including respiratory water loss. The metabolic water gain of 0.23 l ($0.0115 \text{ l km}^{-1} \times 20 \text{ km}$) should be subtracted from this, so in hot conditions net extra water loss due to locomotion (after metabolic water gain of 230 ml) would rise from 0.074 l to 3.13 l per 20 km walked.

Additional water loss due to basal metabolism—not directly associated with walking—is uncertain. A wildebeest grazing in a Kenyan arid equatorial grassland in winter had a daily water turnover/requirement²³ of $54 \text{ ml kg}^{-1} \text{ day}^{-1}$ (5.4% body mass per day), measured on an animal (1° S of the equator)^{38,40}, which is equivalent to, for a 220 kg animal, $220 \times 0.054 = 11.9 \text{ l day}^{-1}$.

These data are supported by measurements of wildebeest water requirements in a temperature-controlled room fluctuating between 22 and 40°C ($10.6 \text{ l per day, } n=3$)⁴⁰. However, during migration and dehydration, animals may use less water for digestion and excretion.

Walking 20 km a day in cool conditions would increase net water loss (by the respiratory route owing to increased ventilation for gas exchange ($0.074 \text{ l H}_2\text{O}$)) by $0.074 / 11.9 = 0.6\%$, thus there is an increase of only 0.6% in water use.

In hot conditions, if 3.13 l of water is evaporated for thermoregulation (the complete heat load), this would rise to $11.9 + 3.13 = 15 \text{ l}$; an increase of $3.13 / 11.9 = 26\%$ in daily water use.

Animals become debilitated through dehydration when they reach around 20% body mass loss, although some animals can survive 30% weight loss²⁴, assuming this weight loss is all water—some will actually be body fat and ingesta. Ruminal contents can contain much water, which may buffer body water to some extent, aiding survival. Therefore, $0.2 \times 220 \text{ kg} = 44 \text{ l}$ of water loss will result in 20% dehydration. At a water loss rate of 11.9 l per day , this would happen in just under four days ($44 / 11.9$). For a wildebeest walking 20 km per day in hot conditions, 20% dehydration would occur in three days ($44 / 15$).

If wildebeest muscle efficiency was the same as the value that we measured for the substantially larger cows (41.8%, Extended Data Table 5) rather than 62.6%, then to generate the same mechanical work (which is converted to heat in the body), the COT and oxygen requirement would increase by $62.6 / 41.8 = 1.50 \times$. This would result in a concomitant increase in ventilation and net water loss by that route of $0.074 \times 1.5 = 0.11 \text{ l day}^{-1}$ in cool conditions. The total heat generated would increase by the same amount ($1.5 \times 7.58 \text{ MJ} = 11.4 \text{ MJ}$). This would require a net thermoregulation evaporation of $11.4 \times 10^6 / 2,257 = 5.1 \text{ l day}^{-1}$

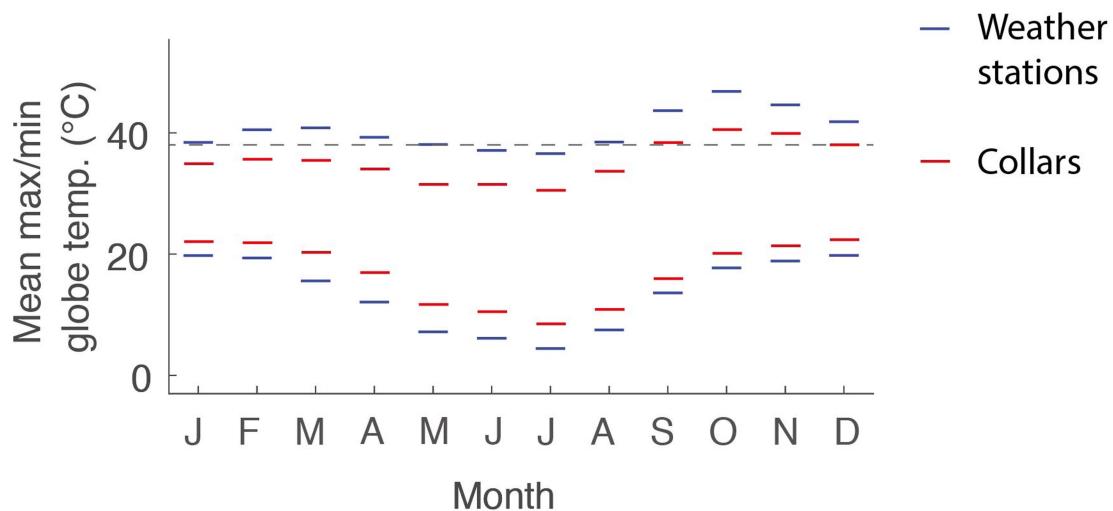
in hot conditions. This would equate to a higher daily water requirement of $11.9 + 5.1 = 17.0 \text{ l day}^{-1}$. This is a 13% rise (17.0/15.0) in daily water requirement.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

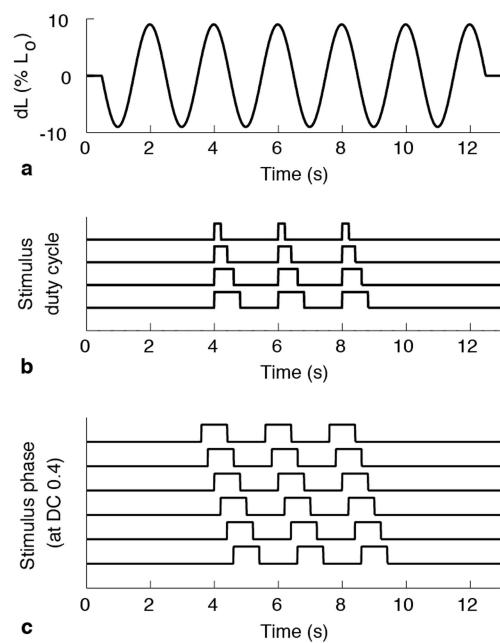
The authors declare that all relevant processed data supporting the findings of this study are available as Source Data. Further data are available from the corresponding authors upon reasonable request.

25. Wilson, A. M. et al. Biomechanics of predator-prey arms race in lion, zebra, cheetah and impala. *Nature* **554**, 183–188 (2018).
26. Dewhurst, O. P. et al. Improving the accuracy of estimates of animal path and travel distance using GPS drift-corrected dead reckoning. *Ecol. Evol.* **6**, 6210–6222 (2016).
27. Hubel, T. Y. et al. Energy cost and return for hunting in African wild dogs and cheetahs. *Nat. Commun.* **7**, 11034 (2016).
28. Hubel, T. Y. et al. Additive opportunistic capture explains group hunting benefits in African wild dogs. *Nat. Commun.* **7**, 11033 (2016).
29. West, T. G. et al. Power output of skinned skeletal muscle fibres from the cheetah (*Acinonyx jubatus*). *J. Exp. Biol.* **216**, 2974–2982 (2013).
30. Barclay, C. J., Woledge, R. C. & Curtin, N. A. Is the efficiency of mammalian (mouse) skeletal muscle temperature dependent? *J. Physiol.* **588**, 3819–3831 (2010).
31. Kretzschmar, K. M. & Wilkie, D. R. The use of the Peltier effect for simple and accurate calibration of thermoelectric devices. *Proc. R. Soc. Lond. B* **190**, 315–321 (1975).
32. Woledge, R. C., Curtin, N. A. & Hornsby, E. Energetic aspects of muscle contraction. *Monogr. Physiol. Soc.* **41**, 1–357 (1985).
33. Hill, A. *Trails and Trials in Physiology* (E. Arnold, London, 1965).
34. Phillips, S. K., Takei, M. & Yamada, K. The time course of phosphate metabolites and intracellular pH using ^{31}P NMR compared to recovery heat in rat soleus muscle. *J. Physiol.* **460**, 693–704 (1993).
35. Curtin, N. A. & Woledge, R. C. Efficiency of energy conversion during sinusoidal movement of white muscle fibres from the dogfish, *Scyliorhinus canicula*. *J. Exp. Biol.* **183**, 137–147 (1993).
36. Barclay, C. J. Mechanical efficiency and fatigue of fast and slow muscles of the mouse. *J. Physiol.* **497**, 781–794 (1996).
37. Curtin, N. A. & Woledge, R. C. Efficiency of energy conversion during shortening of muscle fibres from the dogfish *Scyliorhinus canicula*. *J. Exp. Biol.* **158**, 343–353 (1991).
38. Rumble, J. R. (ed.) *CRC Handbook of Chemistry and Physics* 99th edn (CRC, Boca Raton, 2018).
39. Hetem, R. S. et al. Variation in the daily rhythm of body temperature of free-living Arabian oryx (*Oryx leucoryx*): does water limitation drive heterothermy? *J. Comp. Physiol. B* **180**, 1111–1119 (2010).
40. Maloiy, G. M. O. Water metabolism of East African ruminants in arid and semi-arid regions. *Z. Tierzuecht. Zuchtbiol.* **90**, 219–228 (1973).

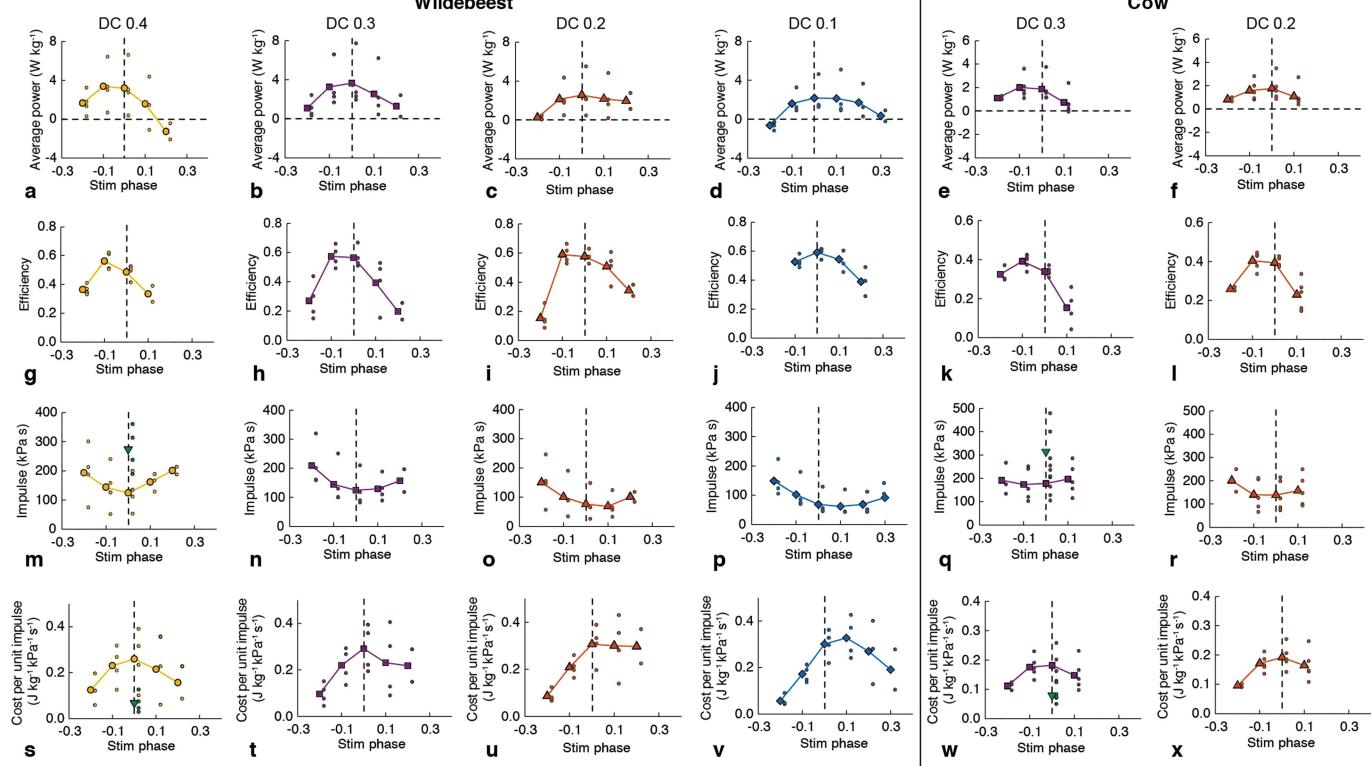


Extended Data Fig. 1 | Comparison of temperature maxima and minima recorded in the collars and at weather stations. The number of working collar sensors varied; therefore, a median was taken from all available data on each day and the maximum and minimum value for each day, which was then averaged over each month. The monthly maximum temperature (mean \pm s.d.) was $5.6 \pm 0.6^\circ\text{C}$ higher at the weather stations than the collars and the monthly minimum temperature was, on average,

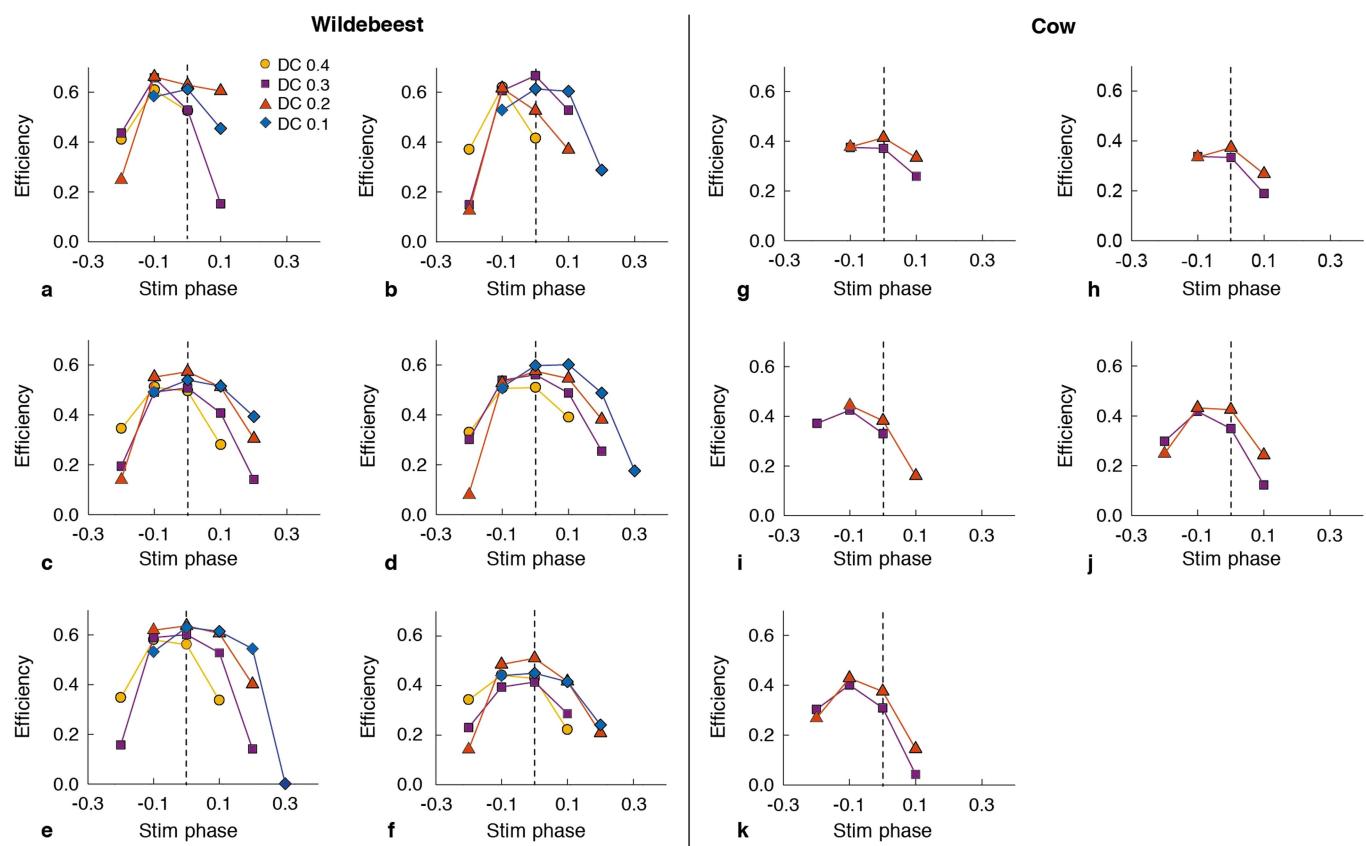
$3.6 \pm 1.1^\circ\text{C}$ lower at the weather stations than the collars. $n = 12$. Ambient temperature exceeded body temperature of 38°C (horizontal dashed line) during nine months of the year. Note weather stations were 10 km away from the river in the dry season range, while animals were in the wet season range to the east from November to April approximately. (Fig. 1b, c).



Extended Data Fig. 2 | Controlled variables for muscle length and stimulation pattern. **a**, Pattern of lever movement. Frequency of 0.5 Hz and peak-to-peak amplitude 18% L_o (10% L_o). L_o is the fibre bundle length at which isometric force was greatest. Values for cow experiments are in parentheses, where they are different than those for wildebeest. **b**, Stimulus duty cycles used in the experiments. Top to bottom, duty cycle of 0.1, 0.2, 0.3 and 0.4 (0.2 and 0.3). **c**, Stimulus phases used in the experiments. Top to bottom: phase $-0.2, -0.1, 0, 0.1, 0.2$ and 0.3 . (-0.2 to 0.1). Phase = 0.0 corresponds to the stimulus starting when shortening starts. In this example, DC = 0.4.



Extended Data Fig. 3 | Individual data points for muscle mechanical and energetic performance. Data presented in Fig. 3, but subdivided by duty cycle. The mean is plotted, symbols and line colours are as in Fig. 3, n numbers are given in Extended Data Table 2.



Extended Data Fig. 4 | Efficiency versus stimulus phase for individual muscle fibre bundles from wildebeest and cows. a–f, Data from wildebeest. g–k, Data from cows. Relationship between stimulus phase and efficiency during three cycles of movement at 0.5 Hz for stimulus duty cycles (DC). Circle, DC = 0.4; square, DC = 0.3; triangle, DC = 0.2; diamond, DC = 0.1. Efficiency = power per rate of heat + work output.

a, b, d, Data are from a different muscle fibre bundle each from a different wildebeest. c, Data are the average of the values shown in e and f, which are results for two fibre bundles from the same wildebeest. g–k, Data are from a different fibre bundle from a different cow.

Extended Data Table 1 | Subject data

Animal name	Start date	End date	Number of days
G690	16-May-2016	27-Oct-2017	529
G692	17-May-2016	10-Apr-2017	328
G693	18-May-2016	17-Nov-2017	548
G694	18-May-2016	16-Nov-2017	547
G695	18-May-2016	14-Nov-2017	545
G696	18-May-2016	14-Nov-2017	545
G697	19-May-2016	18-Nov-2017	548
G698	19-May-2016	22-Aug-2016	95
G699	19-May-2016	19-Nov-2016	184
G700	19-May-2016	31-Aug-2017	469
G701	19-May-2016	13-Nov-2017	543
G702	30-Jul-2016	07-Nov-2017	465
G703	30-Jul-2016	09-Nov-2017	467
G704	20-May-2016	31-Jan-2017	256
G705	30-Jul-2016	09-Nov-2017	467
G706	31-Jul-2016	09-Nov-2017	466
G709	31-Jul-2016	08-Feb-2017	192

Start and end date for data collection and number of days data were collected for each individual. Individual G690 did not migrate but remained in the wet season range throughout the study so only provided data for gait analysis.

Extended Data Table 2 | *n* values

a	DC	0.4	0.3	0.2	0.1
Phase					
	-0.2	4	4	4	4
	-0.1	4	4	4	4
	0	4	4	4	4
	0.1	3	4	4	4
	0.2	2	2	2	3
	0.3				2

b	DC	0.4	0.3	0.2	0.1
Phase					
	-0.2	4	4	4	
	-0.1	4	4	4	4
	0	4	4	4	4
	0.1	2	4	4	4
	0.2		2	2	3

a, c, n values for mean power, impulse and cost per unit impulse in Fig. 3a, c–e, g, h. **a**, Data from wildebeest. **c**, Data from cows. **b, d**, *n* values for mean efficiency in Fig. 3b, f. Note that efficiency was not calculated when net work was negative. **b**, Data from wildebeest. **d**, Data from cows.

c	DC	0.3	0.2
Phase			
	-0.2	3	2
	-0.1	5	5
	0	5	5
	0.1	5	5

d	DC	0.3	0.2
Phase			
	-0.2	3	2
	-0.1	5	5
	0	5	5
	0.1	4	5

Extended Data Table 3 | Minimum cost per unit impulse, values and comparison by species

	Sinusoidal movement	Wildebeest	Cow
		Isometric	Isometric
Mean	0.056	0.068	0.080
Stddev.s	0.022	0.043	0.027
SEM	0.011	0.021	0.012
n	4	4	5

b	Animal	Dur (s)	Initial heat (J/kg)	Impulse (kPa s)	Cost/impulse (J/kg/kPa/s)
	Rat	2	27.1	372	0.073
	Rat	4	51.6	744	0.069
	Wildebeest	2.4	20.4	274.6	0.068
	Cow	2.4	25.0	314.4	0.080

a, Minimum cost per unit impulse ($\text{J kg}^{-1}/(\text{s} \times \text{kPa})$). Mean minimum values (1 value per fibre bundle) for contractions with sinusoidal movement at any of the tested duty cycles and phases (wildebeest) and during isometric contraction at a stimulation duty cycle of 0.4 (wildebeest and cow). **b**, Comparison of cost per unit impulse of muscle fibre bundles from rats, wildebeest and cows. Rat values are based on a previous study³⁵. 'Dur' is the duration of stimulation under isometric conditions. For rat data, isometric stress was 186 kPa (based on wet mass), having been converted from the reported value of 0.93 N m g^{-1} dry mass, using wet mass/dry mass = 5. Impulse was duration \times isometric stress. Wildebeest and cow values are reported here. See text and isometric values (downwards triangles) in Fig. 3d, h. Duration = 3 contractions \times 0.8 s per contraction.

Extended Data Table 4 | Maximum enthalpy efficiency value for muscle fibre bundles from wildebeest and cow

	animal-bundle code	max	net work (J/kg)	enthalpy (J/kg)	avg power (W/kg)	avg enthalpy rate (W/kg)	DC	phase
a								
	Wildebeest							
	1A	0.662	11.76	17.76	1.96	2.96	0.2	-0.1
	2A	0.666	11.57	17.36	1.93	2.89	0.3	0
	3A&B	0.573	13.03	21.58	2.17	3.60	0.2	0
	4A	0.602	30.40	50.52	5.07	8.42	0.1	0.1
	mean	0.626	16.69	26.80	2.78	4.47		
	stdev.s	0.046	9.16	15.93	1.53	2.65		
	sem	0.023	4.58	7.96	0.76	1.33		
	n	4	4	4	4	4		
b								
	animal-bundle code	max	net work (J/kg)	enthalpy (J/kg)	avg power (W/kg)	avg enthalpy rate (W/kg)	DC	phase
	Wildebeest							
	3A	0.636	20.48	32.20	3.41	5.37	0.2	0
	3B	0.510	5.59	10.96	0.93	1.83	0.2	0
c								
	animal-bundle code	max	net work (J/kg)	enthalpy (J/kg)	avg power (W/kg)	avg enthalpy rate (W/kg)	DC	phase
	Cow							
	1	0.415	20.75	50.05	3.46	8.34	0.2	0
	2	0.373	6.47	17.35	1.08	2.89	0.2	0
	3	0.443	4.79	10.80	0.80	1.80	0.2	-0.1
	4	0.432	8.52	19.73	1.42	3.29	0.2	-0.1
	5	0.429	11.93	27.82	1.99	4.64	0.2	-0.1
	mean	0.418	10.49	25.15	1.75	4.19		
	stdev.s	0.027	6.32	15.19	1.05	2.53		
	sem	0.012	2.83	6.79	0.47	1.13		
	n	5	5	5	5	5		

a, b. Data from wildebeest. **c.** Data from cows. Animal-bundle code: digit indicates the animal, letter indicates the fibre bundle from that animal. In **a**, the line 3A&B lists the averages of the results for two fibre bundles, A and B, from the same wildebeest, no. 3. In **b**, the results of these fibre bundles, 3A and 3B, are listed separately. The table lists the maximum enthalpy efficiency (max ϵ), the duty cycle and phase at which it was produced and the net work, enthalpy, power and enthalpy rate produced in the max ϵ condition (maximum ϵ value from all duty cycles and phases tested on the muscle fibre bundle, see Extended Data Fig. 2). Enthalpy efficiency (ϵ) is the net work/enthalpy produced by the muscle during three cycles of sinusoidal movement at 0.5 Hz with stimulation during part of each movement cycle. Net work is the sum of the work done during the shortening part and the lengthening part of the movement cycles. In this context, work done during shortening is taken to be positive, and that during lengthening to be negative. Enthalpy is the sum of net work and heat production. Work is the integral of active force and length change. Active force is the total force produced with stimulation – that produced without stimulation. ‘Av power’ is the average power in three cycles of movement = work/6 s. ‘Av enthalpy rate’ is the enthalpy produced in three cycles of movement/6 s. Duty cycle (stimulation duration in s in one cycle/2-s cycle time) and stimulation phase (that is, the stimulation start time – shortening start time)/2-s cycle time) are shown.

Extended Data Table 5 | Values of maximum enthalpy efficiency for locomotor muscles from different species

animal	max ϵ	(sem; n)	design	move freq (Hz)	muscle	ref
wildebeest	0.626	(\pm 0.023; 4)	C	0.5	flexor carpi ulnaris	this ms
cow	0.418	(\pm 0.012; 5)	C	0.5	flexor carpi ulnaris	this ms
dogfish	0.41	(\pm 0.02; 13)	C	2 & 2.5	white myotomal	Curtin & Woledge (1993)
mouse	0.34	(\pm 0.03; 4)	C	8	extensor digitorum longus	Barclay (1994)
mouse	0.52	(\pm 0.01; 4)	C	3	soleus*	Barclay (1994)
wild rabbit	0.266	(\pm 0.041; 5)	C	1 or 2	extensor digiti V & peroneus longus	in preparation
tortoise	0.77	(\pm 0.02; 8)	I		rectus femoris	Woledge (1968)
mouse	0.26	(\pm 0.01; 5)	I		extensor digitorum longus	Barclay et al (2010)
mouse	0.333	(\pm 0.02; 6)	I		extensor digitorum longus	Barclay (1996)
mouse	0.425	(\pm 0.025; 6)	I		soleus*	Barclay (1996)
dogfish	0.312	(\pm 0.020; 6)	I		white myotomal	Curtin & Woledge (1991)

Measurements were all made on intact fibre bundles. Design C, cyclic movement at the listed frequency and with intermittent stimulation. Design I, isotonic (that is, force-clamped) or isovelocity (that is, velocity-clamped) conditions following a period under isometric (that is, constant length) conditions. In design I, stimulation was continuous, and efficiency was measured only during shortening. Data are from previous studies^{11,17,31,35-37}.

*Antigravity muscle.

Extended Data Table 6 | Calculation of cross-bridge work from enthalpy efficiency

a	term	definition
	ε_{Max}	maximum observed initial enthalpy efficiency (work/(work+heat) or power/(power+heat rate)
	enthalpy	work + heat
	η_{CB}	cross-bridge thermodynamic efficiency
	g	rate of enthalpy output at max ε , expressed relative to the rate of enthalpy output in isometric contraction
	f_A	activation heat rate/isometric heat rate
	$g-f_A$	the non-activation rate of enthalpy output at max ε , expressed relative to the rate of enthalpy output in isometric contraction.
	ΔH_{Pcr}	-34 kJ/mol, molar enthalpy change of phosphocreatine hydrolysis
	ΔG_{ATP}	free energy of ATP hydrolysis for conditions in muscle, -60.5 kJ/mol, 100 zJ per molecule of ATP
	W_{CB}	maximum measured work output per cross-bridge ATP-splitting cycle, units = zJ
	W_{max}	50 zJ, the theoretical maximum cross-bridge work per attachment cycle: area under the cross-bridge force-extension curve; derived from the T2 curve. See Fig. 16 Barclay (2015)
	W_{CB}/W_{max}	fraction of the theoretically maximum cross-bridge work that is actually achieved by the muscle

b	equation	definition	c	Species	ε_{Max}	$Max\eta_{CB}$	W_{CB} (zJ)	W_{CB}/W_{max} (%)
	$Max\eta_{CB}$	$= \varepsilon \times (g/g-f_A) \times (\Delta H_{Pcr}/\Delta G_{ATP})$		Dogfish white	0.33	0.22	21.6	43
	W_{CB}	$= \eta_{CB} \times \Delta G_{ATP}$		Mouse Edl	0.26	0.19	18.5	37
				Tortoise	0.77	0.46	45.8	92

d	Species	ε_{Max}	g	f_A	ΔH_{Pcr}	ΔG_{ATP}	η_{CB}	W_{CB} (zJ)	W_{max} (zJ)	W_{CB}/W_{max} (%)
	Wildebeest	0.626	1.9	0.345	34.0	60.5	0.430	42.99	50	86%
		0.626	2.5	0.345	34.0	60.5	0.408	40.81	50	82%
		0.626	5.30	0.345	34.0	60.5	0.376	37.63	50	75%
		0.626	5.3	0.27	34.0	60.5	0.410	41.01	50	82%
		0.626	2.5	0.27	34.0	60.5	0.394	39.44	50	79%
		0.626	5.30	0.27	34.0	60.5	0.371	37.07	50	74%

e	Species	ε_{Max}	g	f_A	ΔH_{Pcr}	ΔG_{ATP}	η_{CB}	W_{CB} (zJ)	W_{max} (zJ)	W_{CB}/W_{max} (%)
	Cow	0.418	1.9	0.345	34.0	60.5	0.287	28.70	50	57%
		0.418	2.5	0.345	34.0	60.5	0.273	27.25	50	55%
		0.418	5.30	0.345	34.0	60.5	0.251	25.13	50	50%
		0.418	1.9	0.27	34.0	60.5	0.274	27.38	50	55%
		0.418	2.5	0.27	34.0	60.5	0.263	26.34	50	53%
		0.418	5.30	0.27	34.0	60.5	0.248	24.75	50	50%

Data were based on a previous study¹⁹. **a**, Definitions of terms. **b**, Equations. **c**, Values for muscle from three species. $Max\eta_{CB}$ and W_{CB} were calculated from the equations in **b** using values of g and f_A from table 11 of the previous study¹⁹. W_{max} is 50 zJ, the theoretical maximum cross-bridge work per attachment cycle. **d**, Maximum enthalpy efficiency and work values for wildebeest muscle. $\varepsilon_{max} = 0.626$ is the maximum enthalpy efficiency. η_{CB} and W_{CB} were calculated using the equations in **b**. We assumed that our ε_{max} from cyclic movement experiments also applies in isotonic or isovelocity experiments (see supporting evidence in Extended Data Table 5). The values of g and f_A are all combinations of the values for dogfish white fibres, mouse EDL fibres and tortoise rectus femoris muscle (see table 11 of the previous study¹⁹). W_{CB}/W_{max} is the work actually done by the cross-bridge as a percentage of the theoretically maximum cross-bridge work. The highest and lowest values of W_{CB}/W_{max} are indicated in bold. **e**, Values for cow fibre bundles corresponding to those described in **d**.

Dopamine enhances signal-to-noise ratio in cortical-brainstem encoding of aversive stimuli

Caitlin M. Vander Weele^{1,4}, Cody A. Siciliano^{1,4}, Gillian A. Matthews^{1,4}, Praneeth Namburi¹, Ehsan M. Izadmehr¹, Isabella C. Espinel¹, Edward H. Nieh¹, Evelien H. S. Schut^{1,2}, Nancy Padilla-Coreano¹, Anthony Burgos-Robles¹, Chia-Jung Chang¹, Eyal Y. Kimchi¹, Anna Beyeler¹, Romy Wichmann^{1,3}, Craig P. Wildes¹ & Kay M. Tye^{1,3*}

Dopamine modulates medial prefrontal cortex (mPFC) activity to mediate diverse behavioural functions^{1,2}; however, the precise circuit computations remain unknown. One potentially unifying model by which dopamine may underlie a diversity of functions is by modulating the signal-to-noise ratio in subpopulations of mPFC neurons^{3–6}, where neural activity conveying sensory information (signal) is amplified relative to spontaneous firing (noise). Here we demonstrate that dopamine increases the signal-to-noise ratio of responses to aversive stimuli in mPFC neurons projecting to the dorsal periaqueductal grey (dPAG). Using an electrochemical approach, we reveal the precise time course of pinch-evoked dopamine release in the mPFC, and show that mPFC dopamine biases behavioural responses to aversive stimuli. Activation of mPFC-dPAG neurons is sufficient to drive place avoidance and defensive behaviours. mPFC-dPAG neurons display robust shock-induced excitations, as visualized by single-cell, projection-defined microendoscopic calcium imaging. Finally, photostimulation of dopamine terminals in the mPFC reveals an increase in the signal-to-noise ratio in mPFC-dPAG responses to aversive stimuli. Together, these data highlight how dopamine in the mPFC can selectively route sensory information to specific downstream circuits, representing a potential circuit mechanism for valence processing.

Despite the popularity of the signal-to-noise ratio (SNR) model for mPFC dopamine in computational and theoretical neuroscience, the degree to which it translates across brain functions is unknown. Evidence supporting dopamine-mediated SNR modulations have been found in *ex vivo* preparations⁴, and *in vivo* during auditory stimulus discrimination⁷, visual stimulus discrimination⁸ and working memory⁹. As mPFC neurons respond to both rewarding and aversive stimuli^{10,11}, and dopamine neurons in the ventral tegmental area (VTA) that project to the mPFC (VTA^{DA}-mPFC neurons) are uniquely sensitive to aversive stimuli^{12–16}, we proposed that mPFC neurons encoding aversive or rewarding events are differentially modulated by dopamine.

Dopamine release in the mPFC in response to aversive stimuli has previously been observed with direct but slow^{14,17}, or fast but indirect^{12,16} methodologies. Fast-scan cyclic voltammetry (FSCV) offers a direct measurement of catecholamine neurotransmission with precise temporal resolution, but is rarely used outside the striatum owing to difficulty in discriminating between noradrenaline and dopamine¹⁸. Here we investigated the precise time course of dopamine release using FSCV combined with optical and pharmacological approaches to dissect contributions of VTA^{DA} neurons. Electrodes were aimed at deep layers (5–6) of the mPFC, where VTA^{DA} terminals were densest, relative to locus coeruleus (LC) noradrenaline terminals (LC^{NA}) (Fig. 1a, b), and secured in locations detecting stimulated dopamine release (Extended Data Fig. 1). In tyrosine hydroxylase (TH)::Cre rats, which expressed halorhodopsin (NpHR) in a Cre-dependent manner in VTA^{DA} neurons, we performed tail pinches with and without photoinhibition of VTA^{DA} neurons (Fig. 1c). Photoinhibition of VTA^{DA}

neurons attenuated the pinch-induced signals in the mPFC (Fig. 1d, e). Further, in a separate group of rats, pharmacological inactivation of the LC did not affect pinch-evoked catecholamine release in the mPFC (Extended Data Fig. 1). These data suggest that VTA^{DA}-mPFC neurons contributed the bulk of the rapid pinch-evoked catecholaminergic signal.

Next, we explored the causal relationship between VTA^{DA}-mPFC and valence processing by testing whether this circuit component was sufficient to promote aversion. We used TH::Cre rats to express channelrhodopsin-2 (ChR2) in VTA^{DA} neurons, and implanted optical fibres over the mPFC (Fig. 1f). Activation of VTA^{DA}-mPFC terminals had no effect on behaviour in real-time place avoidance (RTPA) or conditioned place aversion (CPA) assays (Extended Data Fig. 2). However, in light of the model for dopamine involvement in enhancing the SNR, we considered whether dopamine might enhance responses to discrete, predictive cues. We trained rats to associate auditory or visual cues (counterbalanced) with either shock or sucrose delivery. Once rats learned to discriminate the cues predicting shock or sucrose by freezing or approaching the sucrose port, respectively (Extended Data Fig. 2), we tested their behavioural responses to the ‘competition’ of simultaneously presented cues (Fig. 1g) driving conflicting motivational outputs¹⁰ (Fig. 1h). Photostimulation of VTA^{DA}-mPFC (using empirically determined optical parameters, Fig. 1i) during the competition trials caused rats expressing ChR2 to spend significantly less time in the sucrose delivery port and more time freezing compared to controls expressing eYFP (Fig. 1j, k). Taken together, these data suggest that dopamine is released in a time-locked manner upon presentation of an aversive stimulus, and that VTA^{DA} in the mPFC biases behavioural responses towards aversion in the face of conflicting motivational drives.

We next sought to identify distinct, anatomically defined subpopulations in the mPFC that might relay information relevant to processing of aversive information. The mPFC has many downstream projection targets, including the periaqueductal grey (PAG) and nucleus accumbens (NAc) (Extended Data Fig. 3). In animal studies, stimulation of the PAG evokes aversive responses, including defensive and attack behaviours^{19–21}. While projections to the dorsal PAG (dPAG) have been explored in the context of social behaviour²², contributions of the mPFC-dPAG circuit to discrete stimulus processing have not yet been evaluated. Owing to its reported role in reward-related processes, we also investigated the mPFC-NAc projection for comparison^{23–25}. Consistent with previous results²², we found that the mPFC-dPAG circuit and mPFC-NAc projections formed anatomically distinct subpopulations (Extended Data Fig. 3).

To target these pathways, ChR2 or eYFP alone was expressed in either mPFC-dPAG or mPFC-NAc neurons (Fig. 2a and Extended Data Fig. 4). Photostimulation of mPFC-NAc neurons did not produce detectable differences in behaviour between ChR2 and eYFP-expressing groups during RTPA or CPA (Extended Data Fig. 4). By contrast,

¹The Picower Institute for Learning and Memory, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. ²Department of Cognitive Neuroscience, Radboudumc Nijmegen, Nijmegen, The Netherlands. ³Present address: Salk Institute for Biological Sciences, La Jolla, CA, USA. ⁴These authors contributed equally: Caitlin M. Vander Weele, Cody A. Siciliano, Gillian A. Matthews. *e-mail: tye@salk.edu

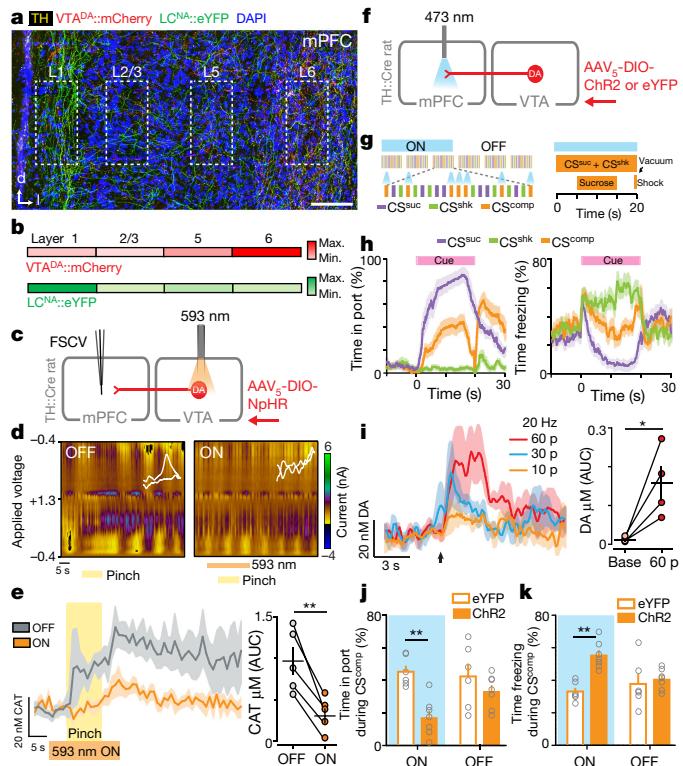


Fig. 1 | Tail pinch evokes rapid dopamine release in the mPFC and dopamine biases behaviour towards aversion during stimulus competition. **a**, Targeted expression of mCherry in VTA^{DA} neurons (VTA^{DA}::mCherry) and eYFP in LC^{NAc} terminals (LC^{NAc}::eYFP) in the mPFC. Scale bar, 100 μ m. **b**, VTA^{DA} terminals were densest in deep layers. LC^{NAc} terminals were densest in superficial layers ($n = 3$ mice). **c**, Strategy to verify dependence of tail-pinch-evoked catecholamine neurotransmission (CAT) on VTA^{DA} neurons. **d**, Representative pseudocolour plots showing tail-pinch-evoked CAT before and during VTA^{DA} inhibition (593-nm laser light, 20 s) ($n = 5$ rats). **e**, Photoinhibition of VTA^{DA} neurons attenuated tail-pinch-evoked CAT release, evident in the average traces (left) and CAT concentration (right). Two-tailed paired t -test, $t_4 = 5.884$, $**P = 0.004$. **f**, Strategy for manipulating dopamine release in the mPFC. **g**, Schematic of competition task. During competition sessions, in addition to sucrose (CS^{suc}, purple) and shock (CS^{shk}, green) trials, sucrose and shock were co-presented as competition trials (CS^{comp}, orange). During ON sessions, VTA^{DA}-mPFC was activated (473 nm, 20 Hz, 60 pulses, every 5 s) during the CS^{comp} trials. During OFF sessions, light was not delivered. **h**, Percentage of time spent in the reward port and freezing during each trial type ($n = 13$ rats). **i**, Evoked dopamine release in the mPFC following 20-Hz optical activation of VTA^{DA} neurons expressing ChR2-mCherry (VTA^{DA}::ChR2-mCherry) ($n = 4$ rats; 60 pulses: two-tailed paired t -test, $t_3 = 3.72$, $*P = 0.034$). Arrow indicates stimulation onset. **j**, Average time spent in the reward port during competition ON trials was lower in VTA^{DA}::ChR2 rats ($n = 7$ rats; closed bars) compared with VTA^{DA}::eYFP control rats ($n = 6$ rats; open bars). Repeated measures, two-way ANOVA, $F_{1,11} = 8.13$, $P = 0.0157$; Bonferroni multiple comparisons tests, $**P = 0.0025$. **k**, Mean time spent freezing during competition ON trials was greater in rats expressing ChR2 compared with rats expressing eYFP. Repeated measures, two-way ANOVA, $F_{1,11} = 13.29$, $P = 0.0039$; Bonferroni multiple comparisons tests, $**P = 0.0013$. Error bars (**e**, **h**, **i**) and shading (**e**, **h**, **i**) represent s.e.m.

activation of ChR2 in mPFC-dPAG neurons reduced the time spent in the light-paired chamber in both RTPA (Fig. 2b) and CPA (Fig. 2c), relative to eYFP controls. In the open-field test, which assays locomotor activity and anxiety-related behaviour, photostimulation of mPFC-dPAG did not affect distance travelled (Fig. 2d) or time spent in the centre of the chamber between ChR2- and eYFP-expressing rats (Extended Data Fig. 5). Strikingly, photostimulation of mPFC-dPAG produced an increase in marble burying and time spent digging (Fig. 2e-h and Supplementary Video 1). The effects in the RTPA and marble-burying assays observed

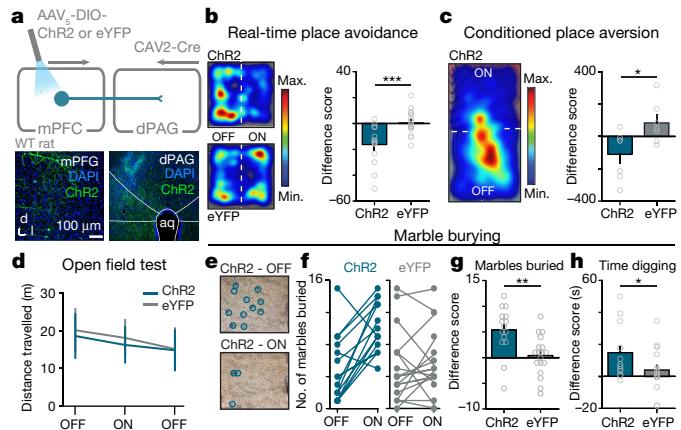


Fig. 2 | The mPFC promotes aversion through projections to the dPAG. **a**, Top, strategy for optogenetic activation of mPFC neurons projecting to the dPAG in wild-type (WT) rats. Bottom, representative images of the mPFC and dPAG. **b**, Left, representative locomotor heat maps of RTPA in rats expressing ChR2 or eYFP in mPFC-dPAG neurons (mPFC-dPAG::ChR2 or mPFC-dPAG::eYFP rats, respectively). Activation of mPFC-dPAG neurons in mPFC-dPAG::ChR2 rats resulted in increased RTPA ($n = 15$ rats) compared to mPFC-dPAG::eYFP controls ($n = 17$ rats) (right). Difference score = (time spent in the ON zone) – (time spent in the OFF zone). Two-tailed unpaired t -test, $t_{30} = 3.902$, $***P = 0.0005$. **c**, Left, representative locomotor heat map of a mPFC-dPAG::ChR2 rat on a CPA test day. Right, activation of mPFC-dPAG neurons resulted in increased CPA in mPFC-dPAG::ChR2 rats ($n = 7$ rats) compared to mPFC-dPAG::eYFP controls ($n = 7$ rats). Two-tailed unpaired t -test, $t_{12} = 2.638$, $*P = 0.0217$. **d**, Optogenetic activation of mPFC-dPAG did not change locomotor activity. mPFC-dPAG::ChR2, $n = 15$ rats; mPFC-dPAG::eYFP, $n = 18$ rats. Distance travelled, two-way repeated measures ANOVA, group \times epoch interaction, $F_{2,62} = 0.94$, $P = 0.3957$. **e**, Representative arena of mPFC-dPAG::ChR2 rat after marble-burying assay when laser stimulation was OFF or ON. **f**, Number of marbles buried in ON and OFF conditions for mPFC-dPAG::ChR2 or mPFC-dPAG::eYFP rats. **g**, Optogenetic stimulation of mPFC-dPAG neurons resulted in a larger change in the number of marbles buried by mPFC-dPAG::ChR2 rats ($n = 15$ rats) compared to mPFC-dPAG::eYFP controls ($n = 18$ rats). Difference score = (number of marbles buried during ON session) – (number of marbles buried during OFF session). Two-tailed unpaired t -test, $t_{31} = 3.341$, $**P = 0.0022$. **h**, mPFC-dPAG::ChR2 rats ($n = 13$ rats) spent more time digging during optical stimulation in comparison to mPFC-dPAG::eYFP controls ($n = 16$ rats). One-tailed unpaired t -test, $t_{27} = 1.961$, $*P = 0.0301$. Data are mean \pm s.e.m.

upon activation of mPFC-dPAG somata were reproduced by activation of mPFC terminals directly in the dPAG (Extended Data Fig. 5).

These data show that optogenetic activation of the mPFC-dPAG projection drives place avoidance and defensive behaviours; however, optogenetic activation may not reflect endogenous circuit function. To address this, we investigated the dynamics of individual neurons in the mPFC-dPAG and mPFC-NAc populations during shock or sucrose presentation. We performed *in vivo* microendoscopic imaging²⁶ of neurons expressing a genetically encoded calcium indicator (GCaMP6m)²⁷. To visualize changes in intracellular calcium concentration indicative of neural activity, we selectively expressed GCaMP6m in mPFC-dPAG and mPFC-NAc neurons (Fig. 3a). Assessment of bulk fluorescence activity, a measure of population activity, revealed that the mPFC-NAc population was not significantly modulated by either shock or sucrose (Fig. 3b). By contrast, mPFC-dPAG neurons showed a robust, time-locked increase in activity in response to shock and a decrease in response to sucrose (Fig. 3c). To assess the activity of individual projection-defined neurons, we used a constrained non-negative matrix factorization algorithm optimized for microendoscopic imaging (CNMF-E)²⁸ (Fig. 3d and Supplementary Videos 2, 3). We identified 169 mPFC-NAc and 118 mPFC-dPAG neurons, which sorted into 6 functional clusters (Fig. 3e and Extended Data Fig. 6). When comparing the normalized responses of individual cells within

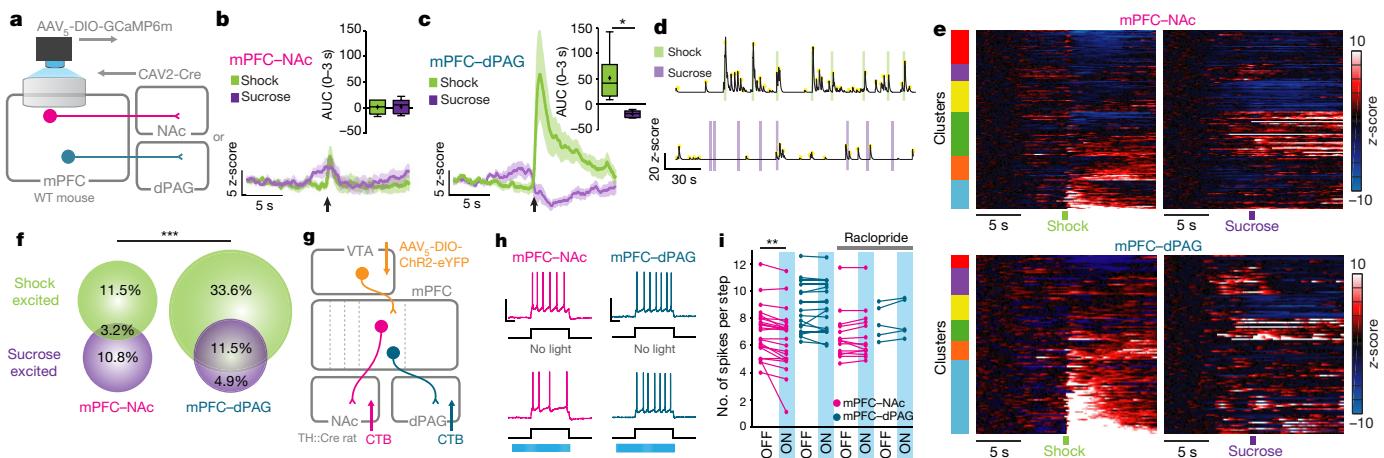


Fig. 3 | mPFC-dPAG neurons preferentially respond to aversive stimuli. **a**, Strategy for recording calcium activity in mPFC-dPAG and mPFC-NAc neurons in wild-type mice. **b**, Bulk fluorescence aligned to shock and sucrose bout in mice expressing GCaMP6m in mPFC-NAc neurons (mPFC-NAc::GCaMP6m mice) ($n = 5$ mice). Responses to sucrose did not differ from responses to shock (0–3 s AUC) in these mice. Two-tailed paired *t*-test, $t_4 = 0.1482$, $P = 0.8893$. AUC, area under the curve. **c**, Bulk fluorescence in mPFC-dPAG::GCaMP6m neurons ($n = 6$ mice). Calcium responses to shock were greater than responses to sucrose (0–3 s AUC). Two-tailed paired *t*-test, $t_5 = 3.743$, $*P = 0.0134$. **d**, Signals were extracted from individual regions of interest (ROIs). Individual transients indicated by yellow dots. **e**, Average traces per ROI aligned to shock or sucrose for each population. Agglomerative clustering results are shown in the bars on the left of each heat map. **f**, The distribution of shock- and sucrose-excited cells for mPFC-dPAG::GCaMP6m ($n = 118$ ROIs) was different from that

for mPFC-NAc::GCaMP6m ($n = 169$ ROIs) ($\chi^2 = 14.76$, $***P = 0.0006$). **g**, Strategy for manipulation of VTA^{DA}-mPFC::ChR2 and recording from dPAG or NAc projectors ex vivo. **h**, Representative traces from mPFC-NAc and mPFC-dPAG neurons during a current step without and with activation of VTA^{DA}-mPFC (470 nm, 20 Hz, 60 pulses). **i**, Optical activation of VTA^{DA}-mPFC did not influence mPFC-dPAG neurons ($n = 17$ cells), but decreased the number of spikes per step in mPFC-NAc neurons ($n = 24$ cells), an effect not observed upon treatment with raclopride (a D2-type dopamine-receptor antagonist) (mPFC-dPAG, $n = 5$ cells; mPFC-NAc, $n = 14$ cells). Two-tailed repeated measures ANOVA, $F_{3,56} = 5.331$, $P = 0.0027$, Bonferroni multiple comparisons tests, mPFC-NAc OFF versus mPFC-NAc ON, $**P < 0.001$. Shading represents s.e.m., boxes show median, first and third quartiles, points indicate the mean and whiskers show minimum and maximum (b, c). Scale bars (electrophysiology): *x* axis, 500 ms; *y* axis, 50 mV.

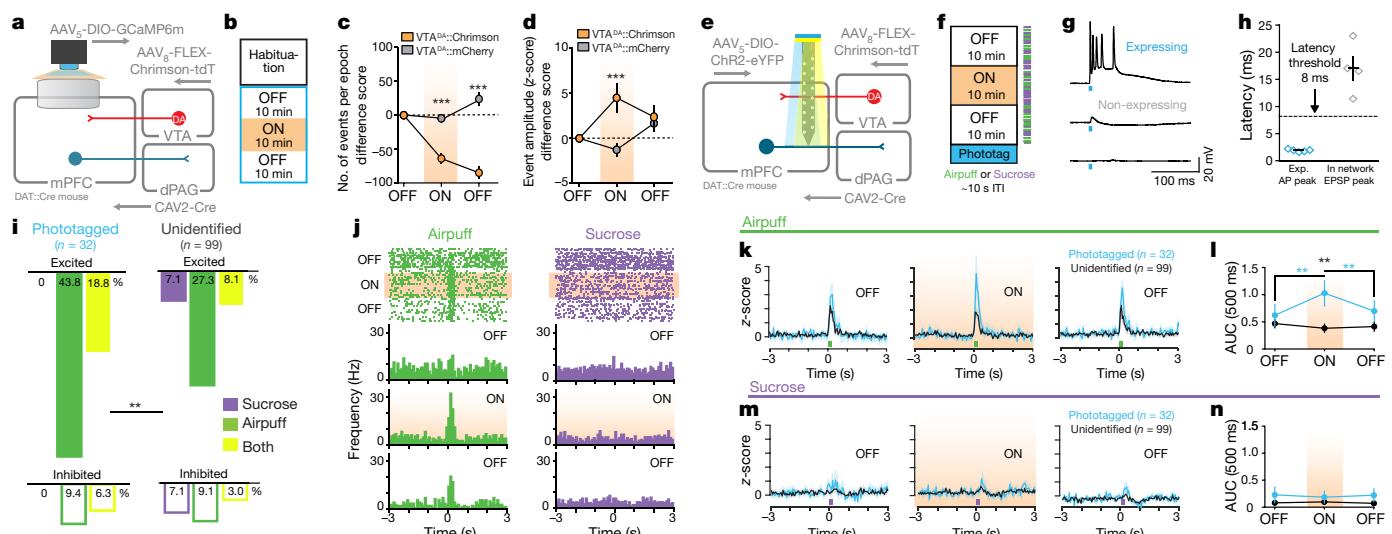


Fig. 4 | Dopamine enhances the SNR of mPFC-dPAG neural responses to aversive stimuli. **a**, Strategy for imaging activity in mPFC-dPAG::GCaMP6m neurons and activation of VTA^{DA}-mPFC in vivo ($n = 3$ mice, 5 recording sessions). **b**, During the ON epoch, VTA^{DA}-mPFC::Chrimson terminals were stimulated with 590-nm light (20 Hz, 60 pulses, every 30 s). **c**, Stimulation of VTA^{DA}-mPFC terminals decreased event frequency (Chrimson, $n = 4$ mice, 44 ROIs; mCherry control, $n = 5$ mice, 50 ROIs). Two-way repeated measure ANOVA, $F_{2,184} = 57.61$, $P < 0.0001$; Bonferroni multiple comparisons tests, $***P < 0.0001$. **d**, VTA^{DA}-mPFC stimulation increased event amplitude. Two-way repeated measure ANOVA, $F_{2,184} = 5.843$, $P = 0.0035$; Bonferroni multiple comparisons tests, $***P < 0.0001$. **e**, Strategy for manipulation of VTA^{DA}-mPFC and identification of mPFC-dPAG::ChR2 using in vivo electrophysiology. **f**, During the ON epoch, VTA^{DA}-mPFC::Chrimson were stimulated with 593-nm light (20 Hz, 60 pulses, every 30 s). Mice received random sucrose and airpuff deliveries. ITI, inter-trial interval. **g**, Representative traces from ChR2-expressing and non-ChR2-expressing

neurons in response to blue light ex vivo. **h**, Latency to action potential (AP) peak for ChR2-expressing ($n = 5$ cells) and to excitatory postsynaptic potential (EPSP) peak for non-ChR2-expressing ($n = 4$ cells) neurons. **i**, Excitatory response patterns were different between populations ($\chi^2 = 9.52$, $P = 0.0016$). **j**, Representative peri-stimulus time histogram (PSTH) of mPFC-dPAG neurons. **k**, Population z-score for phototagged and unidentified units aligned to airpuff. **l**, Stimulation of VTA^{DA}-mPFC neurons enhanced airpuff responses in phototagged, but not unidentified neurons. Two-way repeated measure ANOVA, $F_{2,258} = 6.196$, $P = 0.0024$; Bonferroni multiple comparisons tests, phototagged OFF1 versus ON, $**P = 0.0014$; phototagged ON versus OFF2, $**P = 0.0091$; unidentified OFF1 versus ON versus OFF2, $P > 0.05$; phototagged ON versus unidentified ON, $**P = 0.0012$. **m**, Population z-score for phototagged and unidentified units aligned to sucrose. **n**, VTA^{DA}-mPFC did not change responses to sucrose. Two-way repeated measure ANOVA, $F_{2,258} = 0.4420$, $P = 0.6432$. Error bars (c, d, h, l, n) and shading (k, m) represent s.e.m.

each population, mPFC-NAc responses were heterogeneous while mPFC-dPAG responses were robustly biased towards shock (Fig. 3f and Supplementary Videos 4, 5). Further, transients in mPFC-dPAG neurons were both more frequent and higher in amplitude during shock sessions, compared to those in mPFC-NAc neurons (Extended Data Fig. 6).

On the basis of these functional and anatomical differences, we next sought to assess the impact of dopamine on mPFC-NAc and mPFC-dPAG neurons. To test whether dopamine had different effects on the intrinsic excitability of these populations, we performed whole-cell patch-clamp recordings in acute slice preparations of the mPFC containing VTA^{DA}-mPFC terminals expressing ChR2 and retrogradely labelled mPFC-dPAG or mPFC-NAc neurons (Fig. 3g). We delivered current steps to evoke intermediate levels of neural firing that were paired with photostimulation of VTA^{DA}-mPFC neurons on interleaved sweeps (Fig. 3h). Photostimulation of VTA^{DA}-mPFC neurons reduced the number of spikes per step for mPFC-NAc neurons, but did not detectably alter the excitability of mPFC-dPAG neurons (Fig. 3i). Dopamine-mediated suppression of mPFC-NAc neurons was blocked by the D2-type dopamine receptor antagonist raclopride (Fig. 3i). To investigate dopamine receptor localization on mPFC-NAc and mPFC-dPAG neurons, we performed retrograde labelling of projectors in Drd1a-Cre and Drd2-Cre mice injected with adeno-associated virus (AAV) for Cre-dependent expression of eYFP. We found that mPFC-NAc projectors expressed both D1 and D2 dopamine receptors, whereas mPFC-dPAG projectors largely did not express them (Extended Data Fig. 7). Since dopamine did not modulate mPFC-dPAG neurons ex vivo and this population did not robustly express dopamine receptors, we considered the possibility that dopamine modulates the SNR of incoming sensory information—a function that is only revealed when such inputs are intact.

To investigate this idea, we simultaneously recorded calcium dynamics in mPFC-dPAG neurons while stimulating VTA^{DA} terminals in vivo. Expression of the fluorescent calcium sensor GCaMP6m was targeted to mPFC-dPAG neurons, and dopamine neurons were transduced with the depolarizing red-shifted opsin Chrimson²⁹ or mCherry using a dopamine transporter (DAT)::Cre mouse (Fig. 4a). VTA^{DA}-mPFC terminals were activated during a 10-min ‘laser-ON’ epoch, flanked by two ‘laser-OFF’ epochs without photostimulation (Fig. 4b). Consistent with the model in which dopamine increases the SNR of mPFC-dPAG activity, VTA^{DA}-mPFC stimulation decreased calcium event frequency (Fig. 4c and Extended Data Fig. 8) and increased event amplitude (Fig. 4d and Extended Data Fig. 8). To demonstrate alterations in the SNR, we next explored how dopamine altered activity in mPFC-dPAG neurons in the presence of aversive signals. To test this, we used ChR2-assisted photoidentification of mPFC-dPAG projectors during electrophysiological recordings, coupled with optical manipulations of VTA^{DA}-mPFC (Fig. 4e). VTA^{DA}-mPFC terminals were stimulated during a 10-min ‘laser-ON’ epoch flanked by two ‘laser-OFF’ epochs in an awake, in vivo head-fixed preparation³⁰. During recording, unpredicted sucrose and airpuff presentations were interleaved and mPFC-dPAG neurons expressing ChR2 were optically tagged with blue light at the end of the session (Fig. 4f). Of the 204 total mPFC units recorded, 32 were photoidentified as mPFC-dPAG projectors using an ex vivo verified photoresponse latency threshold (Fig. 4g, h). Consistent with our results from in vivo calcium imaging, a large proportion of mPFC-dPAG neurons were excited by airpuff (Fig. 4i). Stimulation of VTA^{DA}-mPFC terminals did not change basal firing rates in phototagged or unidentified populations (Extended Data Fig. 9). Examination of time-locked neural activity revealed a selective dopamine-mediated amplification of airpuff responses (Fig. 4j–l), but not sucrose responses (Fig. 4m, n) in mPFC-dPAG neurons. This increase in SNR was not observed in the unidentified or photoinhibited populations (Extended Data Fig. 10).

Threatening environmental stimuli require immediate disengagement from ongoing behaviour and engagement of escape and avoidance strategies, which requires tuning of valence-defined circuits. We speculate that dopamine in the mPFC primes top-down

neural circuits that encode aversive stimuli in order to promote avoidance and escape-related defensive behaviours. These findings have clinical relevance to neuropsychiatric disorders characterized by dopamine dysregulation in the mPFC. Our data suggest that mesocortical dopamine governs information routing down discrete mPFC projections and highlights the need for targeted circuit-specific dopamine therapies in the mPFC.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0682-1>.

Received: 14 October 2016; Accepted: 4 September 2018;

Published online: 07 November 2018

1. Arnsten, A. F. T. Stress signalling pathways that impair prefrontal cortex structure and function. *Nat. Rev. Neurosci.* **10**, 410–422 (2009).
2. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).
3. Cohen, J. D., Braver, T. S. & Brown, J. W. Computational perspectives on dopamine function in prefrontal cortex. *Curr. Opin. Neurobiol.* **12**, 223–229 (2002).
4. Kroener, S., Chandler, L. J., Phillips, P. E. M. & Seamans, J. K. Dopamine modulates persistent synaptic activity and enhances the signal-to-noise ratio in the prefrontal cortex. *PLoS ONE* **4**, e6507 (2009).
5. Rolls, E. T., Loh, M., Deco, G. & Winterer, G. Computational models of schizophrenia and dopamine modulation in the prefrontal cortex. *Nat. Rev. Neurosci.* **9**, 696–709 (2008).
6. Winterer, G. & Weinberger, D. R. Genes, dopamine and cortical signal-to-noise ratio in schizophrenia. *Trends Neurosci.* **27**, 683–690 (2004).
7. Popescu, A. T., Zhou, M. R. & Poo, M.-M. Phasic dopamine release in the medial prefrontal cortex enhances stimulus discrimination. *Proc. Natl. Acad. Sci. USA* **113**, E3169–E3176 (2016).
8. Noudoost, B. & Moore, T. Control of visual cortical signals by prefrontal dopamine. *Nature* **474**, 372–375 (2011).
9. Williams, G. V. & Goldman-Rakic, P. S. Modulation of memory fields by dopamine D1 receptors in prefrontal cortex. *Nature* **376**, 572–575 (1995).
10. Burgos-Robles, A. et al. Amygdala inputs to prefrontal cortex guide behavior amid conflicting cues of reward and punishment. *Nat. Neurosci.* **20**, 824–835 (2017).
11. Euston, D. R., Gruber, A. J. & McNaughton, B. L. The role of medial prefrontal cortex in memory and decision making. *Neuron* **76**, 1057–1070 (2012).
12. Kim, C. K. et al. Simultaneous fast measurement of circuit dynamics at multiple sites across the mammalian brain. *Nat. Methods* **13**, 325–328 (2016).
13. Lammel, S., Ion, D. I., Rooper, J. & Malenka, R. C. Projection-specific modulation of dopamine neuron synapses by aversive and rewarding stimuli. *Neuron* **70**, 855–862 (2011).
14. Abercrombie, E. D., Keefe, K. A., DiFrischia, D. S. & Zigmond, M. J. Differential effect of stress on *in vivo* dopamine release in striatum, nucleus accumbens, and medial frontal cortex. *J. Neurochem.* **52**, 1655–1658 (1989).
15. Thierry, A. M., Tassin, J. P., Blanc, G. & Glowinski, J. Selective activation of mesocortical DA system by stress. *Nature* **263**, 242–244 (1976).
16. Mantz, J., Thierry, A. M. & Glowinski, J. Effect of noxious tail pinch on the discharge rate of mesocortical and mesolimbic dopamine neurons: selective activation of the mesocortical system. *Brain Res.* **476**, 377–381 (1989).
17. Finlay, J. M., Zigmond, M. J. & Abercrombie, E. D. Increased dopamine and norepinephrine release in medial prefrontal cortex induced by acute and chronic stress: effects of diazepam. *Neuroscience* **64**, 619–628 (1995).
18. Heien, M. L. A. V., Phillips, P. E. M., Stuber, G. D., Seipel, A. T. & Wightman, R. M. Overoxidation of carbon-fiber microelectrodes enhances dopamine adsorption and increases sensitivity. *Analyst* **128**, 1413–1419 (2003).
19. Bandler, R. & Carrive, P. Integrated defence reaction elicited by excitatory amino acid microinjection in the midbrain periaqueductal grey region of the unrestrained cat. *Brain Res.* **439**, 95–106 (1988).
20. Deng, H., Xiao, X. & Wang, Z. Periaqueductal gray neuronal activities underlie different aspects of defensive behaviors. *J. Neurosci.* **36**, 7580–7588 (2016).
21. Toyote, P. et al. Midbrain circuits for defensive behaviour. *Nature* **534**, 206–212 (2016).
22. Franklin, T. B. et al. Prefrontal cortical control of a brainstem social behavior circuit. *Nat. Neurosci.* **20**, 260–270 (2017).
23. Murugan, M. et al. Combined social and spatial coding in a descending projection from the prefrontal cortex. *Cell* **171**, 1663–1677 (2017).
24. Otis, J. M. et al. Prefrontal cortex output circuits guide reward seeking through divergent cue encoding. *Nature* **543**, 103–107 (2017).
25. Britt, J. P. et al. Synaptic and behavioral profile of multiple glutamatergic inputs to the nucleus accumbens. *Neuron* **76**, 790–803 (2012).
26. Ghosh, K. K. et al. Miniaturized integration of a fluorescence microscope. *Nat. Methods* **8**, 871–878 (2011).
27. Chen, T.-W. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
28. Zhou, P. et al. Efficient and accurate extraction of *in vivo* calcium signals from microendoscopic video data. *eLife* **7**, e28728 (2018).

29. Klapoetke, N. C. et al. Independent optical excitation of distinct neural populations. *Nat. Methods* **11**, 338–346 (2014).
30. Beyeler, A. et al. Divergent routing of positive and negative information from the amygdala during memory retrieval. *Neuron* **90**, 348–361 (2016).

Acknowledgements We thank I. Witten, C. Cameron, N. Parker, M. Murugan, P. Zhou and L. Paninski for advice and code for CNMF-E analysis; M. Schnitzer and D. Cai for advice regarding endoscopic imaging; Y.-N. Leow, A. Shea and N. Golan for histological assistance; N. Imamura and C. Leppla for technical training. We recognize the generosity of the Genetically-Encoded Neuronal Indicator and Effector (GENIE) program, the Janelia Farm Research Campus, V. Jayaraman, R. A. Kerr, D. S. Kim, L. L. Looger and K. Svoboda for providing GCaMP6m. We acknowledge Inscopix for a scientific collaboration and providing early access to nVoke and L. Cardy and A. Stamatakis of Inscopix for technical assistance. We thank E. J. Kremer for providing CAV2-Cre vector; UNC vector core for ChR2, NpHR and ChrimsonR vectors; University of Pennsylvania vector core for GCaMP6m packaging; R. Neve (formerly at the Gene Transfer Core Facility at MIT, now at Massachusetts General Hospital) for packaging the AAV-DIO-synaptophysin-mCherry construct; J. Crittenden for D1-TdTomato/D2-GFP mice and T. Okuyama for Drd1a-Cre and Drd2-Cre mice. K.M.T. is a New York Stem Cell Foundation–Robertson Investigator and a McKnight Scholar, and this work was supported by funding from the JPB Foundation, PIIF, PNDRF, JFDP, Klingenstein Foundation, NARSAD Young Investigator Award, New York Stem Cell Foundation, NIH R01-MH102441-01 (NIMH), NIH Director’s New Innovator Award DP2-DK-102256-01 (NIDDK), and Pioneer Award DP1-AT009925 (NCCIH). C.M.V.W. and E.H.N. were supported by the NSF Graduate Research Fellowship and Integrative Neuronal Systems Training Fellowship (T32 GM007484). C.A.S. is supported by NIH grants F32 MH111216 (NIMH) and K99 DA045103 (NIDA). G.A.M. was supported by

the Charles A. King Postdoctoral Research Fellowship Program, Bank of America, N.A., Co-Trustees. R.W. and N.P.-C. acknowledge funding from the Simons Center Postdoctoral Fellowship. R.W. also recognizes funding from the Netherlands Organization for Scientific Research (NWO) RUBICON. C.A.S., A.B., A.B.-R. and R.W. recognize support from the NARSAD Young Investigator Award.

Reviewer information *Nature* thanks P. Phillips and the anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.M.V.W. and K.M.T. conceived the project. C.M.V.W., C.A.S., G.A.M., E.M.I., I.C.E., E.H.N., E.H.S.S. and N.P.-C. collected data. C.M.V.W., E.H.N., G.A.M., C.A.S., I.C.E., C.-J.C., P.N. and K.M.T. analysed data. E.H.N., P.N., C.-J.C. and E.Y.K. provided MATLAB scripts and advice for data analysis. R.W., A.B., C.P.W. and A.B.-R. provided technical training. C.M.V.W., C.A.S., G.A.M., E.H.N. and K.M.T. contributed to experimental design. C.M.V.W. and K.M.T. wrote the paper. All authors contributed to the editing of the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0682-1>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0682-1>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to K.M.T.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Surgery and viral injections. Specific details of subjects and surgery for each experiment are provided below. All surgery was performed under aseptic conditions and body temperature was maintained with a heating pad. Rodents were anaesthetized with isoflurane mixed with oxygen (5% for induction, 2–2.5% for maintenance, 1 l min^{−1} oxygen flow rate) and placed in a digital small-animal stereotaxic instrument (David Kopf Instruments). Following initial induction, hair was removed from the dorsal surface of the head with hair clippers, ophthalmic ointment was applied to the eyes, the incision area was scrubbed with alcohol pads and betadine (3× each), and 2% lidocaine was injected just under the skin surface above the skull for topical anaesthesia. All measurements were made relative to bregma (unless noted otherwise) for virus/implant surgeries. Viral injections were performed using a bevelled microinjection needle (26 gauge for rats; 33 gauge for mice) with a 10 µl microsyringe (Nanofil; WPI) delivering virus at a rate of 0.05–0.01 µl min^{−1} using a microsyringe pump (UMP3; WPI) and controller (Micro4; WPI). For injections at multiple locations on the dorsal–ventral axis, the most ventral location was completed first and the injection needle was immediately relocated to the more dorsal location for the next injection. After completion of injection, 15 min were allowed to pass before the needle was slowly withdrawn. After viral infusions were completed, craniotomies were filled with bone wax and the incision closed with nylon sutures. Subjects were maintained under a heat lamp and provided 0.05 mg kg^{−1} (rat) or 0.10 mg kg^{−1} (mouse) buprenorphine (subcutaneously, diluted in warm Ringer's solution) until fully recovered from anaesthesia.

All experiments involving the use of animals were in accordance with NIH guidelines and approved by the MIT Institutional Animal Care and Use Committee. For all experiments involving viral or tracer injections, animals containing mistargeted injection(s) were excluded after histological verification.

Viral constructs. Recombinant adeno-associated virus serotype 5 (AAV₅) vectors containing coding sequences for Chr2^{31,32}, NpHR^{33,34}, or fluorescent proteins (mCherry or eYFP) were packaged by the University of North Carolina Vector Core (Chapel Hill, NC). AAV₈-hSyn-FLEX-ChrimsonR-tdTomato²⁹ and AAV₅-hSyn-mCherry were packaged by the University of North Carolina Vector Core (Chapel Hill, NC). Viruses carrying GCaMP6m^{27,35} were packaged by the University of Pennsylvania Vector Core (Philadelphia, PA). Canine adeno-associated virus³⁶ carrying Cre recombinase (CAV2-Cre 4.2 × 10¹² infectious units per ml) was packaged and obtained from E. Kremer (Institut de Génétique Moléculaire de Montpellier, France). AAV₉-hEF1a-DIO-synaptophysin-mCherry was obtained from R. Neve (Viral Gene Transfer Core Facility, MIT, now located at Massachusetts General Hospital).

Catecholamine terminal tracing. Male heterozygous tyrosine hydroxylase (TH)::Cre mice (8–9 weeks old) received unilateral injections of the antero-grade-travelling AAV₅ encoding the fluorescent protein mCherry or eYFP under a double-floxed inverted open-reading frame (DIO) construct (AAV₅-EF1a-DIO-mCherry or AAV₅-EF1a-DIO-eYFP) in the ventral tegmental area (VTA; AP: −3.4, ML: +0.4, DV: −4.25 (1 µl)) and locus coeruleus (LC; AP: −5.45, ML: 1.25, DV: −4.0 and −7.8 (0.5 µl)), counterbalanced. Mice (*n* = 3) were given 10 weeks for viral expression and trafficking of the fluorescent protein to terminals in the medial prefrontal cortex (mPFC). After virus incubation, mice were transcardially perfused, and tissue was sectioned and prepared for immunohistochemistry to label TH⁺ neurons for histological analyses (described below). For quantification of fluorescently labelled TH⁺ neurons in the LC and VTA, single *z*-stacks in the medial ventral tegmental area (VTA) and central LC were acquired using a scanning confocal microscope (Olympus FV1000) with Fluoview software (Olympus) under a 60×/1.42 numerical aperture (NA) oil-immersion objective. The number of co-labelled (TH⁺ and eYFP/mCherry⁺) neurons and eYFP/mCherry-only labelled neurons were counted. *z*-stack stitches encompassing both prelimbic (PL) and infralimbic (IL) regions of the mPFC were acquired under a 40×/1.30-NA oil-immersion objective. Quantification of fluorescence intensity as a proxy for terminal density was accomplished by analysing 100 (w) × 200 (h) µm sections across mPFC layers based on DAPI density/morphology in Fiji³⁷. Sections were normalized to the section with peak fluorescence within subjects. Sample size was based on reports in related literature and was not predetermined by calculation.

Fast-scan cyclic voltammetry. *Subjects.* For FSCV, male and female heterozygous BAC transgenic TH::Cre rats³⁸ (~220 g body weight) were dual housed with ad libitum access to water on a normal 12:12-h light:dark cycle (lights on at 09:00). *Surgery.* TH::Cre rats, which had received a unilateral injection of 2 µl AAV₅ encoding ChR2-mCherry or halorhodopsin 3.0 (NpHR)-eYFP, under a double-floxed inverted open-reading frame construct (AAV₅-EF1a-DIO-ChR2-mCherry or AAV₅-EF1a-DIO-NpHR-eYFP) in the VTA (AP: −5.3, ML: +0.7, DV: −8.2 and −7.8 (1 µl each)) were given at least 8 weeks for viral expression before recording. *In vivo* FSCV experiments were conducted similar to those previously described^{39,40}. Rats were anaesthetized with urethane (1.5 g kg^{−1}, intraperitoneally) diluted in sterile saline and placed in a stereotaxic frame located in a Faraday cage. For both experiments, a glass-encased carbon fibre electrode (~120 µm exposed

carbon fibre, epoxied seal) was lowered into the mPFC (AP: +3.2, ML: +0.8 mm relative to bregma; DV: −2.0 mm from brain surface) through a small craniotomy performed above the deep layers of the mPFC for voltammetric recordings.

For ChR2 experiments (*n* = 5), additional craniotomies were performed above the VTA (AP: −5.5, ML: −0.6 mm), LC (AP: −9.75, ML: −1.25 mm), and contralateral cortex. An Ag/AgCl reference electrode, chlorinated just beforehand, was implanted in the contralateral cortex. A manually constructed optical fibre⁴¹ (400-µm core, 0.48 NA; Thorlabs) cut to 8 mm in length, held in a 2.5 mm ferrule (Precision Fibre Products), was implanted above the VTA (DV: −7.0 mm), and a 26-gauge guide cannula (PlasticsOne) was positioned over the LC (DV: −6.6 mm). Implants were secured to the skull with adhesive cement (C&B Metabond; Parkell).

After the cement dried, the optic fibre implant was connected to a patch cable (Doric) via a ceramic sleeve (PFP) and both reference and carbon-fibre recording electrode were connected to the FSCV interface via a custom-made head stage (S. Ng-Evans, P. E. M. Phillips Laboratory, University of Washington). Dopamine release was evoked by optical activation of the VTA using 150 pulses of 473-nm light (25 mW, 5-ms pulse duration) at 50 or 30 Hz, delivered via a diode-pumped solid state (DPSS) laser (OEM Laser Systems) through the attached patch cable and controlled using a Master-8 pulse stimulator (A.M.P.I.). Electrodes were stereotactically lowered in 0.2-mm increments until optimal dopamine release was detected by photoactivation of VTA dopamine neurons. Optically evoked dopamine release was not detected from one subject for unknown reasons; however, tail-pinch-evoked catecholamine release was observed with characteristic cyclic voltammograms (CVs) for catecholamines, and therefore this rat was included in analyses.

For NpHR experiments (*n* = 5), craniotomies (in addition to that above the mPFC) were performed above the VTA (AP: −5.5, ML: −0.6 mm), nucleus accumbens shell (NAc, AP: 1.5, ML: +0.9), and contralateral cortex. An Ag/AgCl reference electrode, chlorinated just beforehand, was implanted in the contralateral cortex and secured to the skull with adhesive cement (C&B Metabond; Parkell). After the cement dried, reference and carbon-fibre recording electrodes were connected to an FSCV interface via a head stage and the recording electrode was stereotactically lowered into the NAc shell (DV: −6.6 mm relative to brain surface). Following equilibration (see below), a combination bipolar electrical stimulation electrode and 26-gauge guide cannula (PlasticsOne) was stereotactically lowered above the VTA (DV: −6.5 mm) in 0.2-mm increments until dopamine release was detected in the NAc by electrical activation of VTA dopamine neurons via 60 Hz, 60 pulses (biphasic, 200 µA), controlled by an ISO-Flex stimulus isolator (A.M.P.I.). Following dopamine detection, the combination electrical stimulation–guide cannula electrode was cemented in place (C&B Metabond; Parkell) slightly dorsal of the VTA and the carbon-fibre recording electrode transferred into the mPFC (DV: −2.0 mm) and allowed to equilibrate. Sample sizes were based on reports in related literature and were not predetermined by calculation.

Data acquisition. For both experiments, electrodes were allowed to equilibrate for 20 min at 60 Hz and 10 min at 10 Hz. Voltammetric recordings were collected at 10 Hz by applying a triangular waveform (−0.4 V to +1.3 V to −0.4 V, 400 V s^{−1}) to the carbon-fibre electrode versus the Ag/AgCl reference implanted in the contralateral cortex⁴². Data were collected in 60-s files with the tail-pinch onset occurring 10 s into the file for a duration of 10 s (TarHeelCV). Files were collected every 120 s and background subtracted at the lowest current value before pinch onset. Evoked signals maintained characteristic CVs for dopamine and noradrenaline¹⁸, with oxidation and reduction peaks at ~+0.65 V and ~−0.2 V, respectively. For ChR2 experiments, five tail-pinch recordings were obtained with a 120-s inter-recording interval, before LC inactivation. After recordings were completed, 1 µl of tetrodotoxin (TTX, 10 ng per 1.5 µl artificial cerebral spinal fluid) mixed with Fast Green (for spread visualization) was injected into the LC via a microinjection needle through the 26-gauge guide cannula controlled by a syringe pump. Two minutes following infusion completion, five tail-pinch recordings were obtained with a 120-s inter-recording interval, post-LC inactivation. For NpHR experiments, recordings were similarly obtained as 10 recordings at 120-s inter-recording interval. Trials were interleaved with no optical manipulation trials (OFF trials) and trials in which VTA dopamine neurons were inhibited with 20 s constant 598-nm DPSS laser light (5 mW) delivered by a stripped 200 µm core patch cable (Doric) inserted into the combination electrical stimulation/guide cannula located dorsal of the NpHR-expressing VTA dopamine neurons (ON trials). Optical inhibition was initiated 5 s into each ON trial (that is, 5 s before tail-pinch onset) and remained for 20 s (that is, ending 5 s after tail-pinch offset). Oscillatory signals were often observed in the mPFC (however, no such signals were detected in the NAc) and were attenuated by tail pinch and electrical stimulation. Trial averaging alleviated oscillatory interference. Following recording, rats were transcardially perfused and fixed (as described below) and processed using immunohistochemistry for TH immunolabelling to confirm viral expression and implant/recording electrode locations using confocal microscopy. Spread of TTX–Fast Green was recorded during tissue sectioning on a freezing, sliding microtome.

Data analysis. Signals were converted to changes in catecholamine concentration using chemometric, principal component regression, and residual analyses using a custom LabView program (Umich CV, courtesy of R. Keithley), as previously described^{43,44} using in vivo optically and electrically evoked CVs and calibration data obtained from an average of 10 electrodes calibrated in known dopamine concentrations and pH units as previously described¹⁷. For quantification of blue-light-evoked dopamine, AUC was calculated during the 5-s pre-stimulation period, compared to the AUC 5 s following the initiation of 20-Hz, 60 pulses laser light. For quantification of tail-pinch-evoked dopamine, AUC was calculated during the 10 s before pinch onset, compared to the AUC during the 10 s following pinch onset. For comparison of pinch-evoked signals in ON and OFF trials in NpHR experiments, AUC was calculated during the 30-s period following pinch onset.

VTA^{DA}-mPFC behavioural optogenetic experiments. *Subjects.* Male heterozygous BAC transgenic TH::Cre rats³⁸ (~220 g) were dual housed with ad libitum access to water on a normal 12:12 h light:dark cycle (lights on at 09:00). About 1 week following viral injection surgeries, rats were individually housed with restricted food access (~16–20 g chow per day) for ~10 weeks, but retained ad libitum access to water. Sample size was based on reports in related literature and was not predetermined by calculation.

Surgery. TH::Cre rats that had received a unilateral injection of 2 μ l AAV₅-EF1a-DIO-ChR2-eYFP ($n = 6$ –8) or AAV₅-EF1a-DIO-eYFP ($n = 5$ –7) in the VTA (AP: −5.3, ML: +0.7, DV: −8.2 and −7.8 (1 μ l each)) were given at least 12 weeks to ensure Cre-specific viral transduction of ChR2 in VTA^{DA} neurons and protein transport to distal terminals in the mPFC. Following incubation, 20G stainless steel cannulae (PlasticsOne) were bilaterally implanted above the mPFC (AP: +3.2–3.6; ML: ±2.0, DV: −2.8; mm relative to bregma at a 15° angle, bilateral). Guide cannulae were secured to the skull with 2–4 skull screws, a layer of adhesive cement (C&B Metabond; Parkell), followed by black cranioplastic cement (Ortho-Jet; Lang) containing gentamicin antibiotic. The implant was allowed to dry completely before closure of the incision with nylon sutures. Dummies (24G cannulae) were inserted into the guide cannulae to prevent clogging.

General testing procedures. On each test day, a 400- μ m core optical fibre was inserted and attached to the cannulae. Optical fibres extended ~250–500- μ m beyond the cannulae tips. Rats were then transferred to their behavioural apparatus and connected to patch cords connected to dual-rotating commutators for testing. Real-time place preference/aversion and conditioned place preference/aversion assays were identical to those described below. Laser light (473 nm) was delivered through the patch cords at 20 Hz, 60 pulses of 5 ms, every 30 s at 20 mW from optic fibre tip. If an optic fibre broke into a guide cannula or if a guide cannula became clogged, the contralateral guide cannula was used for the remaining experiments. Manipulated hemispheres were counterbalanced.

Real-time place preference/aversion. Individual food-restricted rats were placed in a Plexiglas arena (24 in (l) \times 24 in (w) \times 20 in (h)) and were allowed to move freely between two compartments for 1 h in a dimly lit room containing constant white noise (Marpac Dohm-DS dual speed sound conditioner). Entry into one half of the chamber resulted in photostimulation (VTA^{DA}-mPFC::ChR2/eYFP, unilateral 20 Hz, 60 pulses of 5 ms every 30 s, 20 mW; mPFC-dPAG/NAc::ChR2/eYFP, bilateral 20 Hz 5-ms pulses, 12–15 mW, see below). Stimulation and no-stimulation sides were counterbalanced between animals. Rats were tested on two consecutive days, and on the second day the stimulation side and no stimulation side were reversed. A video camera positioned directly above the arena tracked and recorded movement using EthoVision XT (Noldus). All data presented are tracked from the 'centre' of the subject and time spent in each zone was averaged across the two testing sessions. In between subjects, the behavioural chamber was thoroughly cleaned with 10% glass cleanser diluted in double-distilled water (ddH₂O).

Conditioned place preference/aversion. Individual food-restricted rats were placed in a Plexiglass arena (30 in (l) \times 15 in (w) \times 25 in (h)) divided into two compartments: one with vertical stripes and the other with horizontal stripes. On day 1 (habituation), rats were allowed to move freely between the two compartments for 15 min in a brightly lit room containing constant white noise (Marpac Dohm-DS dual speed sound conditioner). Movement was tracked by an overhead video camera positioned above the arena and time spent in each compartment was calculated using EthoVision XT (Noldus). On days 2 and 3, rats were exposed to conditioning sessions (20 min each, 1 per day) during which they were confined to one side of the chamber and received optical stimulation (VTA^{DA}-mPFC::ChR2/eYFP = unilateral 20 Hz, 60 pulses of 5 ms every 30 s, 20 mW; mPFC-dPAG/NAc::ChR2/eYFP = bilateral 20 Hz 5-ms pulses, 12–15 mW, see below) or no stimulation (counterbalanced for order and side across animals). On day 4 (test), rats were placed in the chamber and allowed to freely explore both compartments in the absence of optical stimulation. Again, movement was tracked by an overhead video camera positioned above the arena using EthoVision XT (Noldus) and a time difference score was calculated by subtracting the time spent in the stimulation-paired compartment on the habituation day from the time spent in the stimulation-paired compartment on the test day (test(time spent in paired side) – habituation(time spent in paired side)).

Stimulus competition task. Training and testing procedures were similar to those previously described¹⁰. Training was performed in standard rat operant chambers (23 \times 30 \times 40 cm; Med Associates) located within sound-attenuating cubicles. Each chamber was equipped with a red house light, speakers for the delivery of tone cues, a sucrose port that was equipped with an infrared beam for the detection of port entries and exits, a syringe pump to deliver sucrose (30% in cage water), two light cues on either side of the sucrose port, and a grid floor for the delivery of electrical shocks. Chambers were wiped down with 70% isopropyl alcohol after each session. Before training, rats were pre-exposed to sucrose in their home cage and were magazine trained in the operant boxes (60 min, 20 sucrose deliveries). The first phase of training consisted of Pavlovian reward conditioning in which rats learned to associate a 20-s conditioned stimulus (CS^{suc}, either a light cue or tone cue (5 kHz, 80 dB), counterbalanced between subjects) with sucrose delivery into the reward port (30% sucrose, 120 μ l per trial). Sucrose was delivered over 10 s during the cue presentation (5–15 s, relative to CS^{suc} onset). ITIs were set to an average of 60 s. If sucrose was not consumed (as detected by the lack of a port entry during the 20-s CS^{suc} presentation), sucrose was immediately removed after cue offset via activation of a vacuum tube located in the sucrose port. Rats were trained on sucrose conditioning for 3 days, with each session comprising 25 trials delivered over ~35 min. The second phase of training consisted of four Pavlovian discrimination sessions where conditioned stimuli predicted sucrose (CS^{suc}) or footshock (CS^{shk}) delivery. During these sessions, the opposite conditioned stimulus (either a light cue or tone cue (5 kHz, 80 dB)) co-terminated with 0.5-s footshock (0.60 mA, 19.5–20 s relative to CS^{shk} onset). CS^{suc} and CS^{shk} cues were counterbalanced and presented in a pseudorandom manner. Each session consisted of 40 total trials (20 of each trial type) with a variable ~60-s ITI. During sucrose conditioning and discrimination sessions, animals were unilaterally connected to a rotating commutator via a dummy patch cord, but no laser light was delivered.

The third phase was the stimulus competition test sessions. Before these sessions, an optical fibre was loaded into a guide cannula, connected to a patch cord, and attached to a rotating commutator, identical to the previous phases. During competition sessions, in addition to CS^{suc} and CS^{shk} trials, competition trials were introduced—in which CS^{suc} and CS^{shk} cues and their respective outcomes were co-presented to evoke conflicting motivation between reward- and fear-associated behaviours. One second before competition trials (CS^{comp}), the 473-nm laser was triggered (20 Hz, 60 pulses of 5 ms every 5 s) for the duration of the 20-s compound cue (4 stimulation trains per competition session). Each competition session consisted of 60 total trials (20 of each trial type) with a variable ~60-s ITI.

Data analysis. Sucrose port entries and exits provided a read-out for reward-related behaviour (based on the percentage of time in the port during each trial type) and were sampled from infrared beam breaks (Med-PC IV, Med Associates). Freezing, defined as the lack of all movement other than respiration, provided a read-out for aversively motivated behaviour. Videos were sampled using side-profiled infrared cameras at 30 frames per second and freezing was quantified using an automated custom MATLAB script that calculated frame-by-frame changes in total pixel intensity as an approximation for animal movement. Frame-by-frame motion values were converted into freezing scores using a binary method relative to a motion threshold. This method produced values which are highly correlated with hand-scored measurements of freezing¹⁰. The time spent in the port was subtracted from the freezing quantification, as animals showed little movement while collecting sucrose.

Retrograde cholera toxin-B tracing. *Rats.* Male wild-type Long–Evans rats (~220 g; Charles River Laboratories) were dual housed on a normal 12:12 h light:dark cycle (lights on at 09:00). Rats were prepared for stereotaxic surgery as described above using the viral infusion parameters also described above (under 'Surgery and viral injections'). In brief, 500 nl cholera toxin subunit B (CTB) conjugated to Alexa Fluor 488, 555 or 647⁴⁵ (0.1%, Molecular Probes) was injected into the dorsal periaqueductal grey (dPAG; AP: −6.6, ML: −0.6; DV: −5.4 mm) and NAc shell (AP: +1.5, ML: +0.95, DV: −7.5 mm) (colour counterbalanced between animals). After 7 days, rats were transcardially perfused and histologically prepared. z-stack stitches encompassing both prelimbic (PL) and intralimbic (IL) regions of the mPFC were acquired using a scanning confocal microscope (Olympus FV1000) with Fluoview software (Olympus) under a 40 \times /1.30-NA oil-immersion objective. Quantification of fluorescence intensity across layers was accomplished by analysing 200 (w) \times 400 (h)- μ m sections encompassing ventral PL/dorsal IL across mPFC layers based on DAPI density/morphology in Fiji. Sections were normalized to the section with peak fluorescence within subjects. For cell quantification, the number of CTB-positive and double-positive neurons was counted in both the IL and PL subregions of the mPFC using FluoView software (Olympus). To examine potential projections from the VTA to the dPAG, 14 VTA sections were immunostained for tyrosine hydroxylase (TH) (see below) and z-stacks were captured under a 40 \times /1.30-NA oil-immersion objective. In each stack, 100 DAPI⁺ cells were identified and the proportions of TH⁺ and CTB⁺ cells were counted. Sample size was based on reports in related literature and was not predetermined by calculation.

Mice. Adult male wild-type C56BL/6 mice (\sim 10 weeks of age; Jackson Laboratory) were prepared similarly to methods described above. In brief, 350 nl CTB conjugated to Alexa Fluor 488, 555 or 647 (Molecular Probes) was injected into the dorsal periaqueductal grey (dPAG; AP: -4.2 , ML: -0.5 ; DV: -2.4 mm) and NAc shell (AP: $+1.0$, ML: $+0.75$, DV: -4.5 mm) (colour counterbalanced between animals). Histological, imaging, and data analyses are the same as previously described.

Projection-specific behavioural optogenetic experiments. mPFC-dPAG and mPFC-NAc subjects. Male wild-type Long-Evans rats (\sim 220 g; Charles River Laboratories) were dual housed on a normal 12:12 h light:dark cycle (lights on at 09:00). About 1 week following viral injection surgeries, rats were individually housed with restricted food access (\sim 16–20 g chow per day) for \sim 10 weeks, but retained ad libitum access to water. Rats were maintained on food restriction unless noted otherwise.

Surgery. For projection-specific targeting for behavioural optogenetics, male wild-type Long-Evans rats were bilaterally injected with 1.2 μ l AAV₅-EF1a-DIO-ChR2(H134R)-eYFP in the mPFC at two locations along the dorsal-ventral axis (0.6 μ l each) (AP: $+3.2$; ML: ± 0.75 ; DV: -3.5 and -2.5 ; mm relative to bregma). To achieve projection-specific recombination, retrogradely travelling CAV2-Cre (4.2×10^{12} infectious units per ml; Institut de Génétique Moléculaire de Montpellier, France) was bilaterally injected (0.6 μ l each) in the dPAG (AP: -6.0 ; ML: ± 0.6 ; DV: -5.2 ; mm relative to bregma (0.4 μ l)), or NAc (AP: $+1.4$; ML: ± 1.0 ; DV: -7.4 ; mm relative to bregma (0.5 μ l)). A subset of mPFC-dPAG rats were co-injected with 0.1 μ l AAV₅-hSyn-mCherry to visualize virus spread. About 7 days following virus surgery, rats were individually housed and placed on food restriction. About 10 weeks later, manually constructed optic fibres (400 μ m core, 0.48 NA) (Thorlabs) held in a 2.5-mm ferrule (Precision Fibre Products) were implanted directly above ChR2/eYFP-expressing mPFC neurons projecting to either the dPAG or NAc for projection-specific manipulations (AP: $+3.2$ – 3.6 ; ML: ± 1.5 ; DV: -2.8 ; mm relative to bregma at a 10° angle, bilateral).

For terminal manipulations, AAV₅-CaMKIIa-ChR2-eYFP was bilaterally injected into the mPFC at two locations along the dorsal-ventral axis (0.6 μ l each) (AP: $+3.2$; ML: ± 0.75 ; DV: -3.5 and -2.5 ; mm relative to bregma). About 7 days following surgery, rats were individually housed and placed on food restriction. About 10 weeks later, manually constructed optic fibres (400- μ m core, 0.48 NA) (Thorlabs, Newton) held in a 2.5 mm ferrule (Precision Fibre Products) were bilaterally implanted directly above the dPAG for mPFC terminal manipulations (AP: -6.6 , ML: ± 1.5 , DV: -4.3 mm relative to bregma at a 10° angle, bilateral). For both experiments, optical fibres were secured to the skull with 2–4 skull screws, a layer of adhesive cement (C&B Metabond; Parkell), followed by black cranioplastic cement (Ortho-Jet; Lang) containing gentamicin antibiotic. The implant was allowed to completely dry before closure of the incision with nylon sutures.

Behavioural testing. Testing was performed at \sim 13 weeks following viral injection and \sim 10 days after optical-fibre implantation to allow sufficient time for transgene expression and tissue recovery. Throughout this period, rats were maintained on food restriction (\sim 16–20 g chow per day). Rats were tested during their light phase (09:30–19:00) under food-deprived conditions. Optic fibre implants were connected to a 200- μ m patch cable (Doric) using a ceramic sleeve (PFP), which connected to a bilateral commutator (rotary joint; Doric) by means of an FC/PC adaptor to allow unrestricted movement. A second patch cable, with an FC/PC connector at either end (Doric), connected the commutator to a 473-nm DPSS laser (OEM Laser Systems). A Master-8 pulse stimulator (A.M.P.I.) was used to control the output of the 473-nm laser, with a light power of \sim 10–15 mW (adjusted to account for optic fibre efficiency). Following each day's experimentation, rats were provided their \sim 16–20 g of standard chow after a variable 0.5–4 h window.

Open-field test. Individual food-restricted rats were placed in a Plexiglass arena (24 (l) \times 24 (w) \times 20 (h) in) and were allowed to move freely within the arena for 9 min with light stimulation occurring during the middle 3 min (3 min OFF, 3 min ON, 3 min OFF design) (mPFC-dPAG/NAc:ChR2/eYFP = bilateral 20 Hz 5-ms pulses, 12–15 mW). The room was brightly lit and contained constant white noise (Marpac Dohm-DS dual speed sound conditioner). A video camera positioned directly above the arena tracked and recorded movement using EthoVision XT (Noldus). In order to assess anxiety-related behaviour, the chamber was divided into a centre (40 \times 40 cm) and periphery region. In between subjects, the behavioural chamber was thoroughly cleaned with 0.03% acetic acid diluted in ddH₂O. All data presented are tracked from the 'centre' of the subject.

Marble burying. Individual food-restricted rats were placed in a standard, rectangular rodent cage (33 (w) \times 40 (l) \times 20 (h) cm) containing \sim 7.5 cm of clean standard bedding and 16 black marbles, which was slightly elevated from the floor (1 m). Sixteen 1.3-cm diameter black marbles were placed on top of the even bedding in a 4×4 array separated from the cage sides by \sim 5 cm. Rats were tested across 2 days for 12 min each, counterbalanced for laser stimulation (mPFC-dPAG:ChR2/eYFP = bilateral 20 Hz 5 ms pulses, 12–15 mW) in a brightly lit room containing constant white noise (Marpac Dohm-DS dual speed sound conditioner). Behaviour was recorded via a video camera positioned directly above the arena

using Ethovision XT (Noldus). Photographs of the behavioural arena before (undisturbed) and after each 12-min session were obtained and marbles that were 100% buried were counted. Time spent digging was scored by two experimenters blind to conditions using ODLog (Macropod). Cage exploration time was obtained by subtracting the time spent of scored behaviours from the total session length. The time spent engaging in each behaviour was quantified by taking the average between the two experimenters. One mPFC-dPAG:ChR2 video was corrupted and was not included in analyses. In between subjects a new cage containing fresh bedding was used and marbles were cleaned with 15% isopropyl alcohol diluted in ddH₂O.

Following the conclusion of the experiments, a subset of rats were stimulated for 5 min in a dark, sound-attenuating room (473 nm, 20 Hz, 20 mW, 5 ms pulses) for c-Fos quantification to verify light-evoked activity in ChR2⁺ mPFC-dPAG neurons. Eighty min later, rats were deeply anaesthetized and transferred to the laboratory and transcardially perfused. Sample size was based on reports in related literature and were not predetermined by calculation.

In vivo epifluorescent calcium imaging. Projection-specific subjects. Male wild-type C57BL/6J mice (\sim 8 weeks old; mPFC-dPAG:GCaMP6m and mPFC-NAc:GCaMP6m) or male DAT::IRES-Cre mice⁴⁶ (\sim 8 weeks old; mPFC-dPAG:GCaMP6m + VTA^{DA}::ChrimsonR or mCherry) were group-housed (2–4 subjects per cage) on a 12:12 h reverse light:dark cycle (lights off at 09:00) before and 4 weeks following initial virus and microendoscope (that is, GRIN lens) implant surgery. Following baseplate adhesion, subjects were individually housed and placed on food restriction (3–6 g normal chow per day) with ad libitum access to water for 3–6 days encompassing testing. Sample sizes were based on reports in related literature and were not predetermined by calculation.

Surgeries. Subjects were prepared for in vivo epifluorescent calcium imaging²⁶ similarly to methods described elsewhere^{47,48}. In brief, to achieve projection-specific imaging, a virus encoding Cre-dependent GCaMP6m (AAV₅-CAG-FLEX-GCaMP6m) was injected into the mPFC (AP: $+1.8$, ML: $+0.3$, DV: -2.75 and -2.4 (300 nl each, bevel facing lateral)) and retrogradely travelling CAV2-Cre (Institut de Génétique Moléculaire de Montpellier, France) was injected into the dPAG ($n = 6$; AP: -4.2 , ML: $+0.5$, DV: -2.4 (350 nl)) or the NAc shell ($n = 5$; AP: $+1.0$, ML: $+0.75$; DV: -4.5 (350 nl)). For manipulation of dopamine terminals in mPFC-dPAG:GCaMP6m + VTA^{DA}::ChrimsonR subjects ($n = 4$), DAT::IRES-Cre mice received 1 μ l AAV₈-hSyn-FLEX-ChrimsonR-tdT in the VTA (AP: -3.4 , ML: $+0.4$, DV: -4.25). Control mice (mPFC-dPAG:GCaMP6m + VTA^{DA}::mCherry; $n = 5$), received 1 μ l AAV₅-EF1a-DIO-mCherry into the VTA using the same coordinates. After virus infusions, the mPFC craniotomy was enlarged to >1 mm in diameter and dura removed with a bent 30 gauge beveled needle, but no tissue was aspirated. A 1 mm diameter, \sim 4 mm length gradient refractive index lens (GRIN lens; GLP-1040, Inscopix) was held by vacuum on the tip of a blunted needle surrounded by plastic tubing for stability and was lowered stereotactically through the craniotomy under constant saline perfusion to minimize tissue/blood desiccation. Lenses were implanted slightly posterior and lateral of the needle track for virus infusions to avoid tissue damage in the imaging plane, and were lowered to locations in the ventral PL/dorsal IL subregion of the mPFC (AP: -1.77 , ML: -0.4 , DV: -2.32 , mm from bregma). Lens implants were secured to the skull with a thin layer of adhesive cement (C&B Metabond; Parkell), followed by black cranioplastic cement (Ortho-Jet; Lang) containing gentamicin antibiotic. Lenses were covered with the top of an eppendorf tube and cemented in place with cranioplastic cement for protection during the virus incubation period (3–4 weeks). The implant was allowed to completely dry before closure of the incision with nylon sutures.

Following virus incubation, mice were again anaesthetized with isoflurane, stereotactically secured, and baseplates (Inscopix) were cemented around the lens to support the connection of the miniaturized microscope for in vivo, freely moving imaging. During this procedure, the protective eppendorf cap and supporting cranioplastic cement were removed using a hand drill. The exposed top of the GRIN lens was scrubbed clean with a cotton-tipped applicator soaked with 15% isopropyl alcohol diluted in ddH₂O. Next, a miniaturized microscope (single channel epifluorescence, 475-nm blue LED, Inscopix) with the baseplate attached was stereotactically positioned over the implanted GRIN lens and adjusted in the DV axis in order to focus on visible landmarks (that is, GCaMP6m-expressing neurons and blood vessels). After the focal plane was identified, the microscope/baseplate was raised by \sim 50 μ m, to account for cement shrinkage, and was subsequently cemented in place with pink dental cement (Stoelting). The microscope was then detached from the baseplates, a final layer of black cranioplastic cement (Ortho-Jet; Lang) was applied to prevent light leak, and the implant was covered with a protective plate (Inscopix) until imaging.

Behavioural sucrose/shock paradigm and data acquisition. Following recovery (\sim 7 days), mice were individually housed and food restricted for 2 days and exposed to 30% sucrose solution (diluted in standard tap/cage H₂O) in the home cage. Food-deprived mice were then trained in operant chambers equipped with sucrose lickometers (Med Associates), with a modified spout that extended into the chamber from the recessed opening, for \sim 60 min while connected to a plas-

tic 'dummy' microscope for training and habituation. All animals readily self-administered sucrose via the lickometer after 2 days of training. On the testing day, food-deprived mice were gently restrained and connected with the miniaturized microscope (single channel epifluorescence, 475-nm blue LED, Inscopix) via the baseplate and secured with a small screw on the baseplate. Mice were allowed to recover from restraint for 10 min before the first session was initiated. Mice were exposed to two 15-min imaging sessions ('sucrose' and 'shock'), counterbalanced and separated by a 15-min intermediate epoch, during which the animal remained in the chamber, but no sucrose or footshocks were administered. During 'sucrose' sessions, food-deprived mice were allowed to self-administer sucrose for 15 min via the lickometer they had been exposed to previously. During 'shock' sessions, mice were exposed to 27 mild electric foot shocks (0.2 mA; 1 s duration; 10–60 s ISI) for 15 min. Grayscale tiff images were collected at 20 frames per second using 20–60% of the miniaturized microscope's LED transmission range (nVista HD V2, Inscopix).

Recording from mPFC-dPAG neurons while manipulating VTA^{DA} terminal activity. Following recovery, DAT::Cre mice were individually housed and food-restricted for 2 days before recording. Before the recording day, food-deprived mice were habituated to handling and the nVoke miniaturized microscope (an integrated imaging and optogenetics system, 455-nm blue GCaMP excitation LED, 590-nm amber optogenetic LED, Inscopix). Twenty-four hours before recordings, mice were habituated in their home cage to a dimly lit recording room containing constant white noise (Marpac Dohm-DS dual speed sound conditioner, Wilmington). On the recording day, mice were attached to the nVoke miniaturized microscope and habituated in their home cage for 15 min. After the 15-min habituation, a 30-min recording session, composed of 10-min OFF-ON-OFF epochs, was initiated. Grayscale images were collected at 10 frames per second using 0.094–0.266 mW mm⁻² (estimated light power based on GRIN lens efficiency) of the miniaturized microscope's 455-nm LED transmission range (nVoke 2.1.5., Inscopix). During the ON epoch, 20 Hz, 60-pulse trains (5 ms each) of 620-nm LED light were initiated every 30 s for the duration of the 10-min epoch.

Image processing. Image processing was accomplished using Mosaic software (v.1.1.2., Inscopix). Raw videos were pre-processed by applying $\times 4$ spatial down-sampling to reduce file size and processing time, and isolated dropped frames were corrected. No temporal downsampling was applied. For sucrose/shock experiments, both recordings per animal (that is, 'sucrose' recording and 'shock' recording) were concatenated to generate a single 30-min video. Lateral movement was corrected for by using a portion of a single reference frame (typically a window surrounding a prominent blood vessel or constellation of bright neurons) as previously described^{26,49}. Images were cropped to remove post-registration borders and sections in which cells were not observed. Two methods were used for ROI identification and single-cell fluorescence trace extraction in order to verify that these processes did not significantly change the pattern of results within our datasets. Both methods are described below in 'CNMF-E analyses' (with and without non-negative constraint on temporal components) and 'non-ROI analyses'. The results from the CNMF-E analyses with non-negative constraint are reported in Figs. 3, 4. The results from the CNMF-E analyses without non-negative constraint and non-ROI analyses are reported in Extended Data Figs. 6, 8.

CNMF-E analyses. After motion correction and cropping, recordings were exported as .tif z-stacks and were downsampled to 10 frames per second. We used a constrained non-negative matrix factorization algorithm optimized for micro-endoscopic imaging (CNMF-E)²⁸ to extract fluorescence traces from ROIs. ROIs were defined by manually selecting seed pixels from peak-to-noise (PNR) graphs of the field of view (FOV)²³. Considering calcium fluctuations can exhibit negative transients, associated with a pause in firing^{13,24}, we also performed analyses in which we did not constrain temporal components to >0 —these data are provided in the extended data figures.

Non-ROI analyses. After motion correction and cropping, recordings were converted to changes in fluorescence (F) compared to background fluorescence (F_0) according to the expression $(F - F_0)/F_0$, using the mean t -projection image of the entire movie as reference (F_0). Calcium signals arising from individual regions of interest (ROIs, that is, cells) were identified using independent and principal component analyses (PCA/ICA), as previously described⁵⁰. Identified PCA/ICA filters were thresholded at their half-max values to define possible ROIs. ROIs were then screened for neuronal morphology and only accepted if the thresholded filters included only one contiguous region with an eccentricity of <0.85 and an area between 30–350 pixels. Accepted ROI filters were merged if their areas overlapped by more than 60% after visual confirmation. The accepted ROI filters were then reapplied to the motion-corrected videos to extract dF/F_0 traces for each ROI. In order to correct for bleaching and possible neuropil contamination of the extracted ROI trace, we correct each ROI tracing using signals from the whole field, using a multiple-step procedure: The full ROI trace and the signals from the whole field were filtered using a 30-s median filter to eliminate the influence of sharp transients or outliers. The influence of the surrounding signals on the ROI trace were quan-

tified using regression ('glmfit' in MATLAB). The resulting regression coefficient was then applied to the original, unfiltered trace to regress out the influence of the non-ROI thresholded field on the ROI trace itself. Multiple background subtraction methods were examined and a non-ROI thresholded approach was implemented because 1) this approach excludes subtraction of prominent processes (that is, dendrites and axons) observed in our dataset, and 2) the reasonable correlation coefficients obtained between individual ROIs are consistent with the range that would be expected based on electrical recordings. To acquire the non-ROI thresholded image for background subtraction, max t -projections of individual recordings were created and thresholded to separate ROIs and their processes from the rest of the FOV. Average signal from the remaining pixels was used as a proxy for the whole-field changes in fluorescence, and regressed from the signal from each ROI.

Data analysis. Individual lick bouts were characterized by lick events detected at the sucrose lickometer and events that were separated by >1 s were identified as an individual lick bout. Calcium signals for the bulk FOV fluorescence and for each ROI were aligned to behavioural events (that is, lick bout initiation and shock). Population z-scores were calculated using the period -10 to -5 s before stimuli onset as baseline. ROIs were classified as being stimulus-excited if the average z-score <1 s after stimulus onset was greater than 3.

For agglomerative clustering, we first concatenated average responses of individual neurons aligned to shocks across trials (expressed as $r(\text{shocks})$, in z-score), and its average response aligned to licks across trials (expressed as $r(\text{licks})$, in z-score), such that each row in the heat map corresponds to one neuron. There were 118 neurons from the PAG and 169 neurons from the NAc in total. Agglomerative hierarchical clustering was applied using Ward's Euclidean linkage, followed by a soft normalization: for each neuron, if its maximum absolute z-score was above 1, its z-score at each frame was divided by its maximum z-score across time. If its maximum absolute z-score was below 1, it remained unchanged. Pairs of neurons that were in close proximity were linked. As they were paired into binary clusters, the newly formed clusters were grouped into larger clusters until a hierarchical tree was formed. A threshold at $0.3 \times \text{max}(\text{linkage})$ was set to prune branches off the bottom of the hierarchical tree, and assign all the neurons below each cut to a single cluster. After clusters were constructed, data from the PAG and the NAc separated to generate their individual heat maps using their original average response profiles (without normalization). For both areas, clusters were sorted in an ascending order on the basis of their third quartile of the response to the shocks. Within each cluster, neurons were also sorted in an ascending order on the basis of their response to the shocks. Different bars on the left side of the heat maps correspond to different clusters. The same colour suggests that they belong to the same cluster from the dendrogram. Calcium event quantifications (number and amplitude) were performed in MiniAnalysis (Synaptosoft) using individual ROI traces from the entire session after conversion to z-score. Baseline from the z transform was computed by thresholding the signal at 20% of the signal amplitude. Calcium events with z-scores <5 or those that did not have a >0.5 AUC were not included in analyses because events of this magnitude were not reliably retain transient, calcium-event characteristics across animals. ROIs that did not contain events meeting event criteria were excluded.

Ex vivo electrophysiology to examine dopamine effects on projector populations. **Subjects.** Male and female heterozygous BAC transgenic TH::Cre rats (~ 220 g; Charles River Laboratories) were dual-housed on a normal 12:12 h light:dark cycle (lights on at 09:00) throughout the duration of experiments. Sample sizes were based on reports in related literature and were not predetermined by calculation. **Surgery.** Rats first received bilateral infusions of AAV₅-EF1a-DIO-ChR2-eYFP, as previously described (see 'FSCV Surgery'). Rats were allowed to recover for virus surgery for an 8–10 week incubation period to ensure Cre-specific viral transduction of ChR2 in VTA^{DA} neurons and protein transport to distal terminals in the mPFC. After incubation, rats received a second surgery to retrogradely label dPAG and NAc shell projectors in the mPFC. CTB injections were performed similarly as previously described ('Retrograde cholera toxin-B tracing'). In brief, rats received bilateral injections of CTB conjugated to Alexa Fluor 488 or 555 (Molecular Probes) into the dPAG (AP: -6.6 , ML: -0.6 , DV: -5.4 mm), the NAc (AP: $+1.5$, ML: $+0.95$, DV: -7.5 mm), or one in each hemisphere (fluorophores were counterbalanced between rats).

Brain slice preparation. Seven days following CTB injections, TH::Cre rats were deeply anaesthetized with sodium pentobarbital (250 mg kg⁻¹; intraperitoneal) and transcardially perfused with 60 ml ice-cold modified artificial cerebrospinal fluid (aCSF) (NaCl 87, KCl 2.5, NaH₂PO₄·H₂O 1.3, MgCl₂·H₂O 7, NaHCO₃ 25, sucrose 75, ascorbate 5, CaCl₂·2H₂O 0.5 (composition in mM) in ddH₂O; osmolarity 322–326 mOsm, pH 7.20–7.30) saturated with carbogen gas (95% oxygen, 5% carbon dioxide). Following decapitation, the brain was rapidly removed from the cranial cavity and coronally dissected (AP: ~ -1.5 mm from bregma). Coronal 300- μ m brain sections were prepared from the anterior portion of the brain containing the mPFC and striatum, using a vibrating microtome (Leica VT1000S,

Leica Microsystems). The posterior portion of the brain was transferred to 4% paraformaldehyde (PFA) dissolved in 1× PBS for fixation and subsequent histological processing (see below in 'Histology'). Brain slices were given at least 1 h to recover in a holding chamber containing aCSF (NaCl 126, KCl 2.5, NaH₂PO₄·H₂O 1.25, MgCl₂·6H₂O 1, NaHCO₃ 26, glucose 10, CaCl₂·H₂O 2.4 (composition in mM); in ddH₂O; osmolarity 298–301 mOsm; pH 7.28–7.32) saturated with carbogen gas at 32 °C before being transferred to the recording chamber for electrophysiological recordings.

Whole-cell patch-clamp recordings. Once in the recording chamber, brain slices were continually perfused with fully oxygenated aCSF at a rate of 2 ml min⁻¹ at 30–32 °C. Neurons were visualized using an upright microscope (Scientifica) equipped with IR-DIC optics and a QImaging Retiga EXi camera (QImaging) through a 40× water-immersion objective. Brief illumination through a 470-nm or 595-nm LED light source (pE-100; CoolLED) was used to identify CTB-488 and CTB-555-expressing mPFC neurons, respectively, before recording. Whole-cell patch-clamp recordings were performed using glass electrodes (resistance 4–6 MΩ) pulled from thin-walled borosilicate glass capillary tubing (World Precision Instruments) on a P-97 horizontal puller (Sutter Instrument) and filled with internal solution containing (in mM) potassium gluconate 125, NaCl 10, HEPES 20, Mg-ATP 3, neurobiotin 0.1% in ddH₂O (osmolarity 287, pH 7.33). For electrophysiological recordings, signals were amplified using a Multiclamp 700B amplifier (Molecular Devices), digitized at 10 kHz using a Digidata 1550 (Molecular Devices), and recorded using Clampex 10.4 software (Molecular Devices). Capacitance, series resistance (R_s) and input resistance (R_{in}) were frequently measured during recordings to monitor cell health, using a 5-mV hyperpolarizing step-in voltage clamp. The resting membrane potential and the current–voltage (I–V) relationship of the neuron were determined in current-clamp mode using incremental 20 pA, 500-ms square current pulses from −120 pA to +260 pA. The instantaneous and steady-state action potential firing frequencies were calculated using the first 100 ms and last 300 ms of the current pulse, respectively.

In order to assess the effect of activating ChR2-expressing VTA (dopamine) terminals on mPFC neuron firing, a square current pulse (2-s duration) was applied in current-clamp mode to elicit stable firing (~2–6 Hz). After 20 s a 20-Hz train of 470-nm light (5-ms pulse duration) was delivered through the 40× objective for 3 s. During the last 2 s of this blue-light train, the same square current pulse was applied to the cell. This protocol was repeated every 50 s and the firing during the current pulses (with and without blue light stimulation) was used for analysis. To determine the effect of VTA (dopamine) terminal stimulation on the rheobase of the neuron, the same protocol was performed, but instead of a square current pulse, a 2-s current ramp was applied to the cell.

The D2-type dopamine-receptor antagonist raclopride was used in a subset of recordings during which a square current pulse was applied with and without optical stimulation of ChR2-expressing VTA (dopamine) terminals. Raclopride (Sigma-Aldrich) was prepared fresh at the start of each recording session and was dissolved in aCSF to give a final concentration of 10 μM. Raclopride was perfused onto the slice for at least 10 min before electrophysiological recordings were commenced.

Analysis of action potential firing was performed offline using Clampfit 10.4 software (Molecular Devices) and passive membrane properties were computed using custom MATLAB software written by P.N. based on MATLAB implementation of the Q method⁵¹.

Immunohistochemistry. Following recording, slices were transferred to 4% PFA solution overnight at 4 °C, and were then washed four times (for 10 min each) in 1× PBS. Slices were then blocked in 1× PBS solution containing 0.3% Triton X-100 and 5% normal donkey serum (NDS; Jackson ImmunoResearch) for 1 h at room temperature. They were then incubated in primary antibody solution containing chicken anti-TH antibody (1:1,000; Millipore, MA, USA) in 1× PBS with 0.3% Triton X-100 (Thermo Fisher Scientific) and 3% NDS overnight at 4 °C. Slices were subsequently washed four times (for 10 min each) in 1× PBS and then incubated in secondary antibody solution containing Alexa Fluor 647-conjugated donkey anti-chicken (1:1,000; Jackson ImmunoResearch) and Alexa Fluor 405-conjugated streptavidin (1:1,000; Biotium) in 1× PBS with 0.1% Triton X-100 and 3% NDS for 2 h at room temperature. Slices were finally washed five times (for 10 min each) in 1× PBS, then mounted onto glass slides and coverslipped using polyvinyl alcohol (PVA) mounting medium with DABCO (Sigma-Aldrich).

Ex vivo electrophysiology to determine latency for phototagging experiments.

Subjects and surgery. To verify the latency of blue-light-evoked action potentials in ChR2-expressing mPFC-dPAG projectors, DAT::Cre mice were used, which had received the same viral surgery as those for in vivo electrophysiology experiments. Viral incubation for ex vivo recordings was matched for those for in vivo experiments. For subject and surgery details, see below, 'In vivo electrophysiology, surgery'.

Brain slice preparation. Brain slice preparation was similar to the previously described method in 'ex vivo electrophysiology to examine dopamine effects on projector populations'. In brief, mice were deeply anaesthetized with sodium

pentobarbital (90 mg kg⁻¹; intraperitoneal) and transcardially perfused with 20 ml ice-cold modified aCSF (NaCl 87, KCl 2.5, NaH₂PO₄·H₂O 1.3, MgCl₂·6H₂O 7, NaHCO₃ 25, sucrose 75, ascorbate 5, CaCl₂·2H₂O 0.5 (composition in mM) in ddH₂O; osmolarity 322–326 mOsm, pH 7.20–7.30) saturated with carbogen gas (95% oxygen, 5% carbon dioxide). Following decapitation, the brain was rapidly removed from the cranial cavity and coronally dissected (AP: ~0 mm from bregma). Coronal 300-μm brain sections were prepared from the anterior portion of the brain containing the mPFC and striatum, using a vibrating microtome (Leica VT1000S, Leica Microsystems). The posterior portion of the brain was transferred to 4% PFA dissolved in 1× PBS for fixation and subsequent histological processing (see below in 'Histology'). Brain slices were given at least 1 h to recover in a holding chamber containing aCSF (NaCl 126, KCl 2.5, NaH₂PO₄·H₂O 1.25, MgCl₂·6H₂O 1, NaHCO₃ 26, glucose 10, CaCl₂·H₂O 2.4 (composition in mM); in ddH₂O; osmolarity 298–301 mOsm; pH 7.28–7.32) saturated with carbogen gas at 32 °C before being transferred to the recording chamber for electrophysiological recordings.

Whole-cell patch-clamp recordings. Recordings were similar to those previously described above. In brief, recordings were made from visually identified neurons expressing ChR2-eYFP and non-expressing neighbours. Blue light was provided by a 470-nm LED light source (pE-100; CoolLED) delivered through a 40× immersion objective. ChR2 expression in recorded neurons was confirmed by the presence of sustained inward current in response to 1-s constant pulse of blue light delivered in voltage-clamp mode.

Offline analysis was performed in Clampfit 10.4 software (Molecular Devices). Latency to action potential or excitatory postsynaptic potential (EPSP) peak were averaged from 30 responses to a 5-ms pulse of blue light (delivered in a 10-pulse, 1-Hz train every 60 s). Latency was measured as the duration from the onset of the light pulse to the peak of the action potential or EPSP.

Dopamine receptor localization on projector populations. *Subjects.* Transgenic male and female Drd1a-Cre ($n = 3$, B6.FVB(Cg)-Tg(Drd1a-cre)FK150Gsat/MmuCud; ID# 036916-UCD from MMRRC originally from GENSAT BAC Tg project) and Drd2-Cre mice ($n = 3$, B6.FVB(Cg)-Tg(Drd2-cre)ER44Gsat/MmuCud; ID# 032108-UCD from MMRRC originally from GENSAT BAC Tg project) (~12 weeks old) were group-housed (2–4 subjects per cage) on a 12:12 h reverse light:dark cycle (lights off at 09:00) throughout the duration of experiments with ad libitum access to food and water. Sample sizes were based on reports in related literature and were not predetermined by calculation.

Surgeries. To label Drd1a- and Drd2-expressing mPFC neurons, AAV₅-EF1a-DIO-eYFP was injected bilaterally into the mPFC (AP: +1.8, ML: +0.3, DV: −2.75 and −2.4 (300 nl each, bevel facing lateral)). Mice were allowed to recover and incubate for 4 weeks. In a second surgery, 350 nl CTB conjugated to Alexa Fluor 555, or 647 (Molecular Probes) was injected into the dPAG (AP: −4.2, ML: −0.5; DV: −2.4 mm) and NAc shell (AP: +1.0, ML: +0.75, DV: −4.5 mm) (in contralateral hemispheres, colour counterbalanced) to retrogradely label mPFC-dPAG and mPFC-NAc projectors. Mice were euthanized 6 days later as previously described. Histological, imaging, and data analyses are similar to those described above.

In vivo electrophysiology. *Subjects.* Male DAT::IRES-Cre mice (~6–8 weeks old) were group-housed (2–4 subjects per cage) on a 12:12 h reverse light:dark cycle (lights off at 09:00) throughout the duration of experiments. Two days after head-bar adhesion (~2 weeks before recordings), cages were placed on food restriction (4 h access to standard chow per day) with ad libitum access to water throughout training and recording. Sample sizes were based on reports in related literature and were not predetermined by calculation.

Surgery. To achieve projection-specific ChR2 expression for in vivo photoidentification of mPFC-dPAG projectors, a virus encoding Cre-dependent ChR2 (AAV₅-EF1a-DIO-ChR2-eYFP) was injected into the mPFC (AP: +1.8, ML: +0.3, DV: −2.75 and −2.4 (300 nl each, bevel facing lateral)) and retrogradely travelling CAV2-Cre (Institut de Génétique Moléculaire de Montpellier, France) was injected into the dPAG (AP: −4.2, ML: +0.5, DV: −2.4 (350 nl)). For manipulation of dopamine terminals, DAT::IRES-Cre mice received 1 μl AAV₈-hSyn-FLEX-ChrimsonR-tdT in the VTA (AP: −3.4, ML: +0.4, DV: −4.25).

Head-bar adhesion. After 11+ weeks of virus incubation, and ~2 weeks before behavioural training, mice were briefly anaesthetized and a small aluminium head-bar (2 cm × 2 mm × 2 mm) was placed on the skull 5 mm posterior to the bregma along with one reference and one ground pin contacting the dura mater just anterior to the head-bar, in the contralateral cortex. A small pilot hole was made with a cranial drill above the mPFC and was marked with a pen. The area surrounding the pilot hole/mark was covered with petroleum jelly to prevent covering with dental cement. The three elements (head-bar, ground pin and reference pin) were cemented using one layer of adhesive cement (C&B metabond; Parkell) followed by a layer of cranioplasty cement (Dental cement; Stoelting). After the cement dried, the pilot hole/mark was covered with a silicone gel (Kwik-Sil Adhesive, WPI) to keep the bone clear during behavioural training.

Behaviour. Two days after head-bar adhesion, mice were food restricted and pre-exposed to a 30% sucrose solution. Mice were head-fixed³⁰ in front of two

small tubes, one located just under the nose and the other above it pointed at the nose. The bottom tube delivered sucrose (training and recording days) and the top tube delivered airpuff (recording days only). Mice were trained to retrieve small drops (3 μ l) of sucrose delivered through the bottom tube via a solenoid valve (Parker), measured by breaks of an infrared beam recorded by an Arduino board (SmartProjects). Training sessions gradually increased in total duration (0.5–1.5 h) and sucrose ITIs increased (15–80 \pm 8 s) over 5–8 days. The solenoid valves were triggered with a custom software written in LabVIEW (National Instruments) powered by NIDAQ-6251 and Arduino hardware.

Pre-recording craniotomy. After 5–8 days of habituation and training, mice were briefly anaesthetized with isoflurane (5% for induction, 1.5% after) and placed in a stereotaxic frame while their body temperature was controlled with a heating pad. A craniotomy was performed over the mPFC using the pilot hole/mark previously implemented using a hand-held drill. When the craniotomy was open, the dura was removed, blood cleaned with perfusion of saline, and then covered with petroleum jelly. Mice were removed from the stereotaxic frame and placed in a clean cage while their body temperature was maintained using a heat lamp until they fully recovered from anaesthesia.

In vivo electrophysiological recordings and phototagging. Once the mice recovered from the craniotomy surgery (at least 1 h), they were head-fixed and a silicon optrode (A1x16-Poly2-5mm-50 s-177, NeuroNexus) coated with red fluorescent latex microspheres (Lumafluor) was inserted into the anterior mPFC and lowered from the surface of the cortex for 1 mm at 10 μ m s⁻¹ using a motorized actuator (Z825B - 25 mm Motorized Actuator, Thorlabs) mounted on a shuttle (460A linear stage, Newport) fixed to the stereotaxic arm. Next, the optrode was lowered for 1 mm at 1–2 μ m s⁻¹. During the insertion of the electrode, sucrose was delivered every 60 \pm 8 s. After the probe was lowered to \sim 2 mm below brain surface, sucrose deliveries were halted and a 10-min wait period commenced to let the tissue stabilize around the recording probe. Recording sessions were initiated using a RZ5D TDT system (Tucker-Davis Technologies) while presenting \sim 40 sucrose and 40 airpuff trials (11 \pm 5 s ITI) randomly intermixed throughout the entire 30-min recording period. The recording period was broken into three 10-min epochs: 10 min into the recording period (first OFF epoch), 593-nm laser-light pulse trains (20 Hz, 60 pulses of 5 ms) were delivered through the optrode every 30 s for 10 min (20 pulse trains total, ON epoch). Ten more minutes were recorded in the absence of laser manipulation (second OFF epoch)—resulting in an OFF-ON-OFF epoch structure, with laser delivery occurring only during the ON epoch. Following completion of a 30-min recording session, a photoidentification session using a 473 and/or 405-nm laser was conducted, during which pseudorandomly dispersed stimulations were delivered: 1-s constant light, 10 \times 1 Hz, 5-ms pulse trains, and 100 ms of 100 Hz (5-ms pulses). Recordings were then terminated and the optrode was lowered 300 μ m to a new recording site at 1–2 μ m s⁻¹. The recording protocol was then repeated after a 30-min inter-session interval. Recordings sessions continued until we reached the bottom of the mPFC (\sim 3 mm from brain surface) or when mice became sated and stopped retrieving sucrose. The electrode was then retracted at 5 μ m s⁻¹, the craniotomy cleaned with saline, and covered with silicone gel (Kwik-Sil Adhesive, WPI) to protect the brain until the next day of recording. During the second day of recording, the same procedure was repeated in a more posterior recording location. Following completion of the second day of recordings, mice were anaesthetized with sodium pentobarbital and transcardially perfused. The brain was extracted, sectioned, and examined under a confocal microscope to verify the viral expression and the locations of the recording electrode.

Analysis of in vivo electrophysiological recordings. Recording sessions were exported from the TDT format to Plexon offline sorter using OpenBridge (Tucker-Davis Technologies). Offline sorter (Plexon) was used to sort single units. Neural responses to sucrose/airpuff delivery and light stimulation were visualized through peristimulus time histograms (PSTH) and rasters for every unit using NeuroExplorer.

Data from Plexon and Neuroexplorer data files were then imported into MATLAB and analysed using software written by P.N. Sucrose and airpuff PSTHs for each epoch (OFF-ON-OFF) were *z*-transformed using the histogram values in a 2-s baseline period starting 3 s before the onset of the stimulus. Similarly, PSTHs around a light pulse (used for photoidentification of dPAG projectors) were *z*-transformed using a baseline window of 40 ms before the onset of the light pulse. To test the significance of neural responses, Wilcoxon signed-rank tests were performed on the neural activity of each unit by comparing the number of spikes in a baseline window and an experimental window starting at the onset of stimulus or light pulse. The experimental window for AUC stimulus response was set to 0.5 s. The experimental window for light response was 8 ms based on the results of ex vivo recordings. The significance threshold for the Wilcoxon signed-rank test was set at $P < 0.01$. Latency to the light pulse was defined as the first bin in the PSTH to cross 4 standard deviations relative to the 40-ms baseline window. Only units which met both criteria were considered phototagged and thus mPFC-dPAG projectors. Burst analyses were performed in NeuroExplorer using interval

specifications. Bursts defined as three or more consecutive spikes with an interval of less than 25 ms in between the first two spikes and less than 50 ms in subsequent spikes, as previously defined for the mPFC⁵².

Histology. *Perfusion and storage.* Subjects were deeply anaesthetized with sodium pentobarbital (200 mg kg⁻¹; intraperitoneal injection) and transcardially perfused with 15 ml (mouse) or 60 ml (rat) of Ringer's solution followed by 15 ml (mouse) or 60 ml (rat) of cold 4% PFA dissolved in 1 \times PBS. Animals were decapitated and the brain was extracted from the cranial cavity and placed in 4% PFA solution and stored at 4°C for at least 48 h. Thirty-six hours before tissue sectioning, brains were transferred to 30% sucrose solution dissolved in 1 \times PBS at room temperature. Upon sinking, brains were sectioned at 60 μ m on a freezing sliding microtome (HM420; Thermo Fisher Scientific). Sections were stored in 1 \times PBS at 4°C until immunohistochemical processing.

Immunohistochemistry. Sections were blocked in 1 \times PBS with 0.3% Triton containing 3% NDS (Jackson ImmunoResearch), for 1 h at room temperature followed by incubation in primary antibody solution: chicken anti-TH (1:1,000; Millipore) or rabbit anti c-Fos (1:500; Santa Cruz Biotechnology) in 1 \times PBS with 0.1% Triton containing 3% NDS for 48 h at 4°C. Sections were then washed 4 times (10 min each) with 1 \times PBS and immediately transferred to secondary antibody solution: AlexaFluor 647-conjugated donkey anti-chicken (1:1,000; Jackson ImmunoResearch) or Cy3 donkey anti-rabbit (1:500, Jackson ImmunoResearch) and a DNA-specific fluorescent probe (DAPI; 1:50,000) in 1 \times PBS containing 3% NDS for 2 h at room temperature. Sections not processed for immunohistochemistry were incubated in 1 \times PBS with 0.3% Triton containing 3% NDS and DAPI (1:50,000) for 1 h. Sections were washed 4 times (10 min each) in 1 \times PBS and mounted onto glass slides. Slices were allowed to dry and were coverslipped using PVA mounting medium with DABCO (Sigma). Stereotaxic coordinates were determined using brain atlases for rat⁵³ and mouse⁵⁴.

Confocal microscopy. Fluorescent images were captured using a confocal laser scanning microscope (Olympus FV1000), with Fluoview software (Olympus), under a dry 10 \times / 0.40-NA objective, a 60 \times / 1.42-NA oil-immersion objective, or a 40 \times / 1.30-NA oil-immersion objective. The locations of opsin expression, injection site, lesion from the optic fibre placement, and the position of carbon-fibre recording electrodes were determined by taking serial *z*-stack images through the 10 \times objective across a depth of 20–40 μ m, with an optical slice thickness of 5–8 μ m. High-magnification images for fluorescence quantifications were obtained through the 40 \times or 60 \times objective using serial *z*-stack images with an optical slice thickness of 3–4 μ m (5 slices) using matched parameters and imaging locations. Fluorescence (in arbitrary units) was obtained from analysis using Fiji. For quantitation of fluorescence across layers in the mPFC, measurements were normalized to the *z* stack containing the maximum value.

Sholl analysis. Neurobiotin-filled, streptavidin-stained mPFC-dPAG and mPFC-NAc projectors from ex vivo electrophysiology experiments were imaged at 40 \times (1.30-NA oil-immersion objective) using a confocal laser-scanning microscope (Olympus FV100) covering the whole dendritic and axonal arborization in the slice. Neurons were reconstructed and Sholl analysis (number of intersections, 20- μ m rings from soma) performed using the 'simple neurite tracer' plugin in Fiji (<http://snyderlab.com/2016/05/25/tracing-neurons-using-fiji-image/>).

Statistics. Statistical analyses were performed using GraphPad Prism (GraphPad Software) and MATLAB (MathWorks). Group comparisons were made using one-way or two-way ANOVA followed by Bonferroni post-hoc tests to control for multiple comparisons. Paired and unpaired two-way Student's *t*-tests were used to make single-variable comparisons. Unpaired one-way *t*-tests were used to make comparisons with a priori hypotheses (time spent digging in marble burying assay). Tests for binomial distribution were also used on single populations. Non-parametric Wilcoxon signed-rank tests were used to make comparisons between non-parametric data. χ^2 tests were used to compare distribution of responsive cells between mPFC-dPAG and mPFC-NAc. All statistical tests were two-tailed unless otherwise noted as an a priori hypothesis. Thresholds for significance were placed at * $P < 0.05$, ** $P < 0.01$ and *** $P < 0.001$. All data are shown as mean \pm s.e.m.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Data, unprocessed and unpseudocoloured images, and custom MATLAB codes are available upon request.

31. Nagel, G. et al. Channelrhodopsin-2, a directly light-gated cation-selective membrane channel. *Proc. Natl Acad. Sci. USA* **100**, 13940–13945 (2003).
32. Zhang, F., Wang, L.-P., Boyden, E. S. & Deisseroth, K. Channelrhodopsin-2 and optical control of excitable cells. *Nat. Methods* **3**, 785–792 (2006).
33. Grdinariu, V., Thompson, K. R. & Deisseroth, K. eNpHR: a *Natronomonas* halorhodopsin enhanced for optogenetic applications. *Brain Cell Biol.* **36**, 129–139 (2008).

34. Schobert, B. & Lanyi, J. K. Halorhodopsin is a light-driven chloride pump. *J. Biol. Chem.* **257**, 10306–10313 (1982).

35. Akerboom, J. et al. Genetically encoded calcium indicators for multi-color neural activity imaging and combination with optogenetics. *Front. Mol. Neurosci.* **6**, 2 (2013).

36. Kremer, E. J., Boutin, S., Chillon, M. & Danos, O. Canine adenovirus vectors: an alternative for adenovirus-mediated gene transfer. *J. Virol.* **74**, 505–512 (2000).

37. Schindelin, J., Arganda-Carreras, I., & Frise, E. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

38. Witten, I. B. et al. Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement. *Neuron* **72**, 721–733 (2011).

39. Tsai, H.-C. et al. Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science* **324**, 1080–1084 (2009).

40. Nieh, E. H. et al. Inhibitory input from the lateral hypothalamus to the ventral tegmental area disinhibits dopamine neurons and promotes behavioral activation. *Neuron* **90**, 1286–1298 (2016).

41. Sparta, D. R. et al. Construction of implantable optical fibers for long-term optogenetic manipulation of neural circuits. *Nat. Protoc.* **7**, 12–23 (2011).

42. Robinson, D. L., Venton, B. J., Heien, M. L. A. V. & Wightman, R. M. Detecting subsecond dopamine release with fast-scan cyclic voltammetry *in vivo*. *Clin. Chem.* **49**, 1763–1773 (2003).

43. Keithley, R. B. & Wightman, R. M. Assessing principal component regression prediction of neurochemicals detected with fast-scan cyclic voltammetry. *ACS Chem. Neurosci.* **2**, 514–525 (2011).

44. Keithley, R. B., Heien, M. L. & Wightman, R. M. Multivariate concentration determination using principal component regression with residual analysis. *Trends Anal. Chem.* **28**, 1127–1136 (2009).

45. Conte, W. L., Karnishina, H. & Reep, R. L. Multiple neuroanatomical tract-tracing using fluorescent Alexa Fluor conjugates of cholera toxin subunit B in rats. *Nat. Protoc.* **4**, 1157–1166 (2009).

46. Zhuang, X., Masson, J., Gingrich, J. A., Rayport, S. & Hen, R. Targeted gene expression in dopamine and serotonin neurons of the mouse brain. *J. Neurosci. Methods* **143**, 27–32 (2005).

47. Resendez, S. L. et al. Visualization of cortical, subcortical and deep brain neural circuit dynamics during naturalistic mammalian behavior with head-mounted microscopes and chronically implanted lenses. *Nat. Protoc.* **11**, 566–597 (2016).

48. Jennings, J. H. et al. Visualizing hypothalamic network dynamics for appetitive and consummatory behaviors. *Cell* **160**, 516–527 (2015).

49. Ziv, Y. et al. Long-term dynamics of CA1 hippocampal place codes. *Nat. Neurosci.* **16**, 264–266 (2013).

50. Mukamel, E. A., Nimmerjahn, A. & Schnitzer, M. J. Automated analysis of cellular signals from large-scale calcium imaging data. *Neuron* **63**, 747–760 (2009).

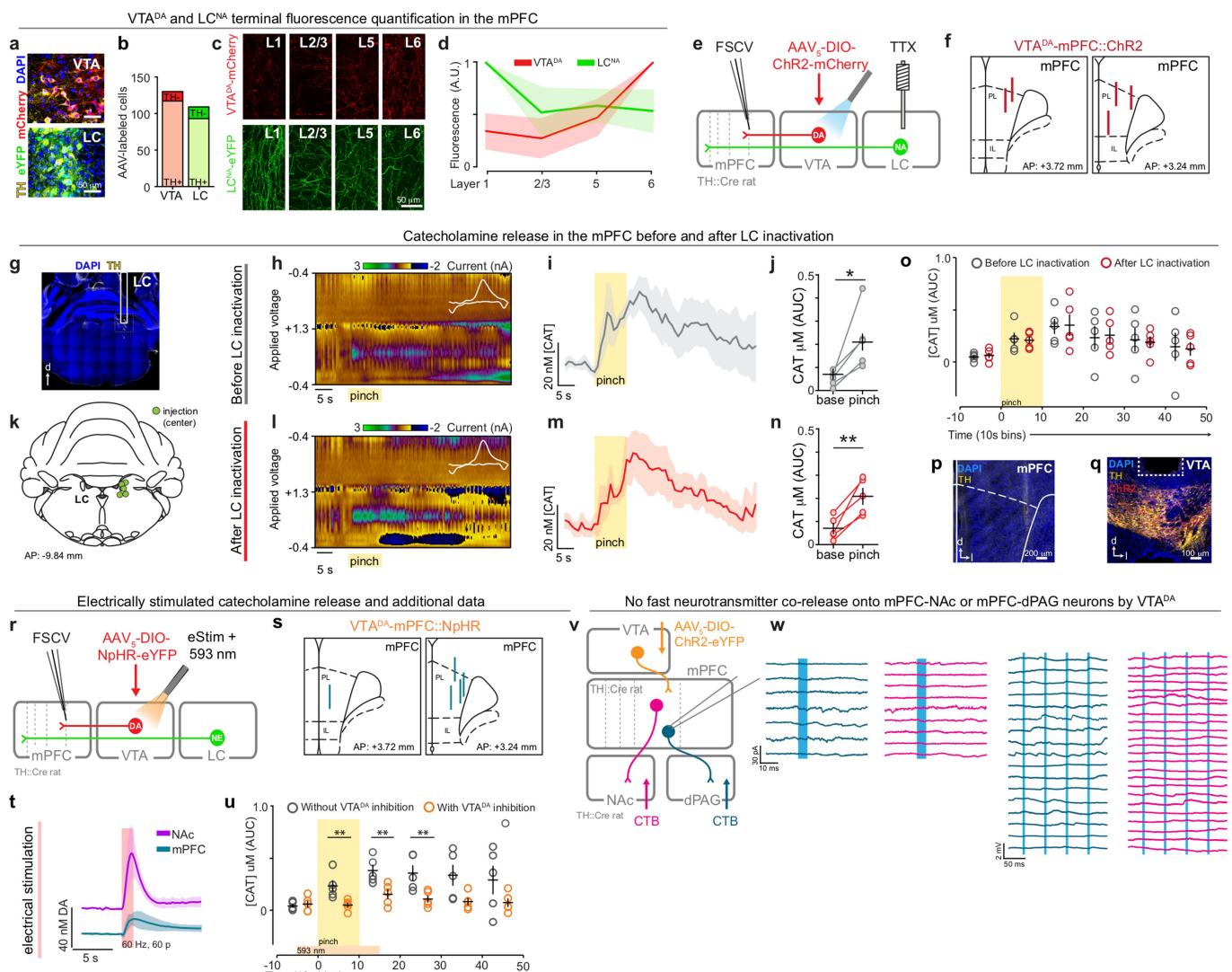
51. Novák, P. & Záhradník, I. Q-method for high-resolution, whole-cell patch-clamp impedance measurements using square wave stimulation. *Ann. Biomed. Eng.* **34**, 1201–1212 (2006).

52. Burgos-Robles, A., Vidal-Gonzalez, I., Santini, E. & Quirk, G. J. Consolidation of fear extinction requires NMDA receptor-dependent bursting in the ventromedial prefrontal cortex. *Neuron* **53**, 871–880 (2007).

53. Paxinos, G. & Watson, C. *The Rat Brain in Stereotaxic Coordinates: Hard Cover Edition* (Academic, Cambridge, MA, 2006).

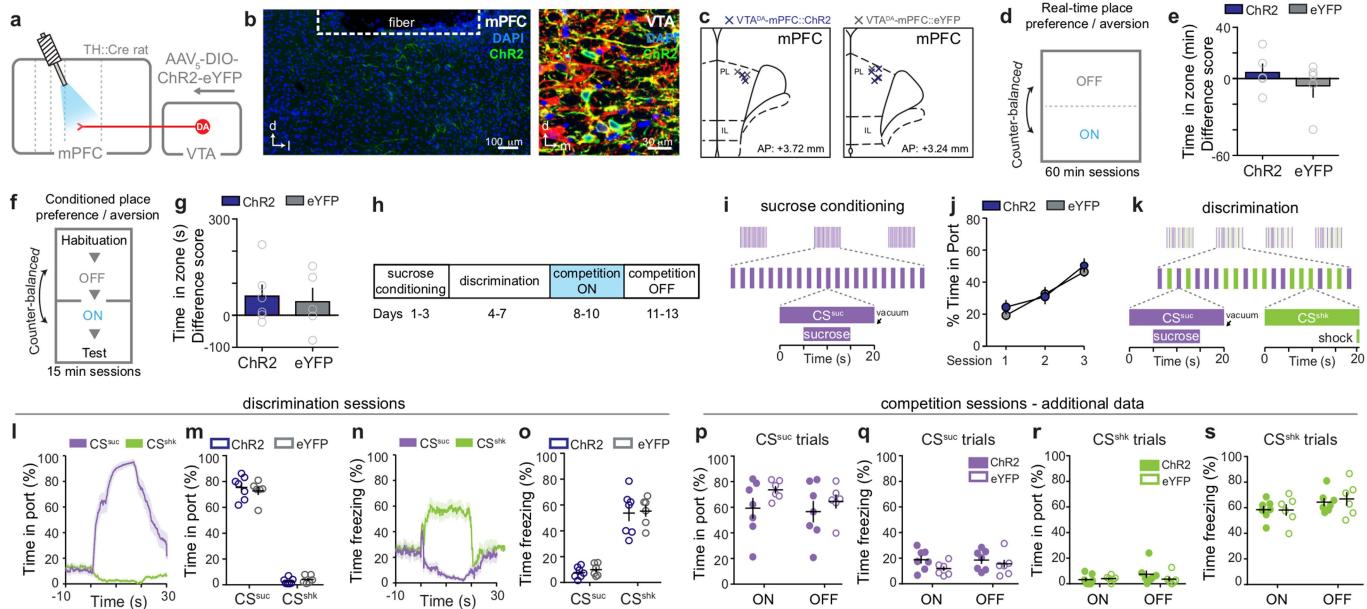
54. Paxinos, G. & Franklin, K. B. *The Mouse Brain in Stereotaxic Coordinates* (Gulf, Houston, 2004).

55. Kupferschmidt, D. A., Juczewski, K., Johnson, K. A. & Lovinger, D. M. Parallel, but dissociable, processing in discrete corticostriatal inputs encodes skill learning. *Neuron* **96**, 476–489 (2017).



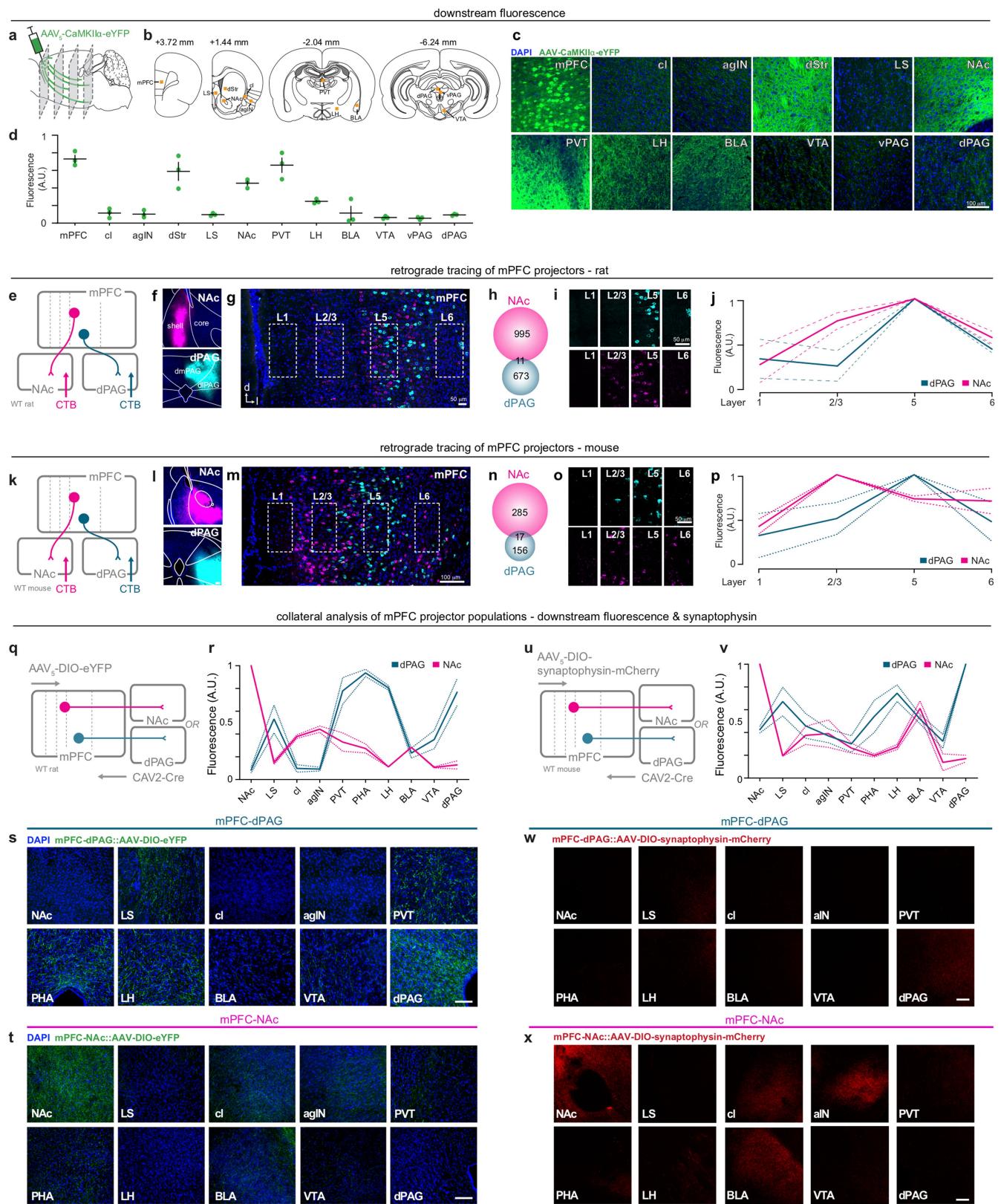
Extended Data Fig. 1 | Investigation of catecholamine terminal density and dopamine release dynamics in the mPFC. **a, b,** Injection of viral constructs (a) enabling Cre-dependent expression into the LC and VTA of TH::Cre mice resulted in (b) fluorescent labelling of TH positive (TH⁺) noradrenergic (NE) neurons in the LC and dopamine neurons in the VTA. **c,** Examination of VTA^{DA} and LC^{NA} fluorescent terminal labelling in the mPFC revealed different patterns of innervation by VTA^{DA} and LC^{NA} neurons across cortical layers in the prelimbic subregion of the mPFC ($n = 3$ mice). **d,** VTA^{DA} terminals were densest in the deep (5 and 6) layers of the mPFC, whereas LC^{NA} terminals were denser in superficial (1 and 2/3) layers. **e,** Schematic of strategy for differentiating dopamine and noradrenaline neurotransmission in the mPFC using FSCV. VTA^{DA} neurons were selectively transduced with ChR2 in TH::Cre rats. After incubation, rats were prepared for anaesthetized FSCV recordings, in which an optical fibre was implanted over the VTA and a guide cannula was positioned over the LC for TTX-mediated pharmacological inhibition. **f,** A glass-encased carbon-fibre recording electrode was lowered into the mPFC for FSCV neurochemical measurements. Schematic representation of all recording electrode locations for ChR2 FSCV experiments. **g,** Representative image of guide cannula track positioned over LC^{NA} cell bodies. Yellow, TH. **h–j,** When VTA^{DA} and LC^{NA} neurons were intact, tail pinch (10 s in duration) rapidly increased extracellular catecholamine concentration (CAT), as shown in a representative false colour plot (h), average CAT trace (i), and concentration quantification (j). $n = 5$ rats; two-tailed paired *t*-test, $t_4 = 3.402$, $*P = 0.027$. Colour plot insets, representative cyclic voltammograms. **k,** TTX-Fast Green injection

locations. **l–n,** After LC inactivation via intra-LC infusion of TTX, tail-pinch-evoked responses were maintained (two-tailed paired *t*-test, $t_4 = 5.249$, $**P = 0.006$. **o,** Pharmacological inactivation of the LC with TTX did not significantly alter tail-pinch-evoked catecholamine release in the mPFC. Two-way repeated measures ANOVA, $F_{5,40} = 0.061$, $P = 0.997$. **p,** Representative image of FSCV electrode track in the mPFC. **q,** Representative confocal image of ChR2-mCherry expression (red) in VTA^{DA} cell bodies. Yellow, TH immunostaining. **r,** Schematic of strategy to verify dependence of pinch-evoked increases in CAT neurotransmission on VTA^{DA} neurons. **s,** Histologically verified FSCV recording electrode locations for NpHR experiments. **t,** Electrical stimulation (60 Hz, 60 pulses, 200 μ A) of the dorsal VTA evoked distinct patterns of dopamine release in the NAc and mPFC ($n = 5$ rats). **u,** Optical inhibition (20 s constant 593 nm, 5 mW) of NpHR-expressing VTA^{DA} neurons attenuated tail-pinch-evoked CAT release in the mPFC. Two-way repeated measures ANOVA, $F_{5,40} = 2.857$, $P = 0.027$; Bonferroni post-hoc tests, $**P < 0.01$. **v,** Schematic of viral strategy to optically manipulate ChR2-expressing VTA^{DA} terminals in the mPFC and record from dPAG- and NAc-projectors retrogradely labelled with CTB with ex vivo electrophysiology. **w,** No evidence of co-release of fast-synaptic neurotransmitters (that is, glutamate and GABA (γ -aminobutyric acid)) from VTA^{DA} terminals onto either mPFC-dPAG (teal) or mPFC-NAc (pink) populations following optical stimulation in voltage-clamp (left) and current-clamp (right). Error bars and shading represent s.e.m. A.U., arbitrary fluorescence units. The rat brain in this figure was reproduced with permission from Paxinos and Watson, 2006⁵³.



Extended Data Fig. 2 | Activation of VTA^{DA} terminals in the mPFC does not support real-time or conditioned place preference. **a**, Schematic of strategy for manipulating dopamine release in the mPFC. VTA^{DA} neurons were selectively transduced with ChR2 in TH::Cre rats and guide cannulae were implanted over the mPFC for the insertion of an optical fibre for light delivery. **b**, Representative confocal image of ChR2-eYFP expression in VTA^{DA}-mPFC underneath a guide cannula (left) and expression in the VTA (right). **c**, Histological verification of guide cannulae placements in the mPFC for ChR2 subjects (left) and eYFP controls (right). **d**, Schematic of experimental design for RTTP/A. When rats entered the ON zone, laser light stimulation was activated for the duration of the time spent in the ON zone (20 Hz, 60 pulses, every 30 s, 20 mW of 473 nm). When rats entered the OFF zone, light stimulation was terminated for the duration of time spent in the OFF zone. **e**, Optogenetic stimulation of VTA^{DA} terminals did not evoke real-time place avoidance or preference in VTA^{DA}-mPFC::ChR2 rats ($n = 5$), compared to VTA^{DA}-mPFC::eYFP controls ($n = 5$), measured by difference score (minutes spent in the ON zone – minutes spent in the OFF zone). Two-tailed unpaired t -test, $t_8 = 0.9337$, $P = 0.3778$. **f**, Schematic of experimental design for CPP/A. Day 1 consisted of a habituation period in which time spent on each compartment of the arena was recorded. On days 2 and 3, a divider was placed in the middle of the chamber to separate the two compartments and rats received either no stimulation (OFF) or stimulation (ON) (20 Hz, 60 pulses, every 30 s, 20 mW), counterbalanced across days. On day 4, the divider was removed and time spent in each compartment was recorded in the absence of stimulation (that is, test day). **g**, Optogenetic stimulation of VTA^{DA} terminals did not support conditioned place aversion or preference in VTA^{DA}-mPFC::ChR2 animals ($n = 6$), compared to VTA^{DA}-mPFC::eYFP controls ($n = 5$), measured by difference score. Two-tailed unpaired t -test, $t_9 = 0.3192$, $P = 0.7569$. **h**, Schematic of task used to examine dopamine modulation of reward and fear-motivated behaviours during competition. **i**, During sucrose training, a conditioned stimulus (CS) (light or tone, counterbalanced) predicted sucrose delivery (CS^{suc}). **j**, During sucrose conditioning, sucrose was delivered to the reward port (vacuum) during CS^{suc} presentation. **k**, During discrimination, the alternative CS (light or tone, counterbalanced) was introduced and predicted foot shock (CS^{shk}). **l**, Average traces showing time spent in the sucrose port before, during, and after each CS presentation (grouped, $n = 13$ rats). **m**, Time spent in the sucrose port did not differ between VTA^{DA}-mPFC::ChR2 rats ($n = 7$) and VTA^{DA}-mPFC::eYFP controls ($n = 6$) during CS^{suc} or CS^{shk} presentation. Repeated measures two-way ANOVA, $F_{1,11} = 0.54$, $P = 0.4789$. **n**, Average traces showing time spent freezing before, during, and after each CS presentation (grouped, $n = 13$ rats). **o**, Time spent freezing did not differ between VTA^{DA}-mPFC::ChR2 rats and VTA^{DA}-mPFC::eYFP controls during CS^{suc} or CS^{shk} presentation. Repeated measures two-way ANOVA, $F_{1,11} = 0.01$, $P = 0.9281$. **p**, During competition sessions, the average time spent in the reward port for CS^{suc} trials during ON sessions and CS^{suc} trials during OFF sessions did not differ between ChR2 rats ($n = 7$, closed) and eYFP controls ($n = 6$, open). Repeated measures two-way ANOVA, $F_{1,11} = 0.82$, $P = 0.3845$. Note that during ON sessions, stimulation was only delivered during the CS^{comp} trials. **q**, Average time spent freezing for CS^{suc} trials during ON sessions and CS^{suc} trials during OFF sessions did not differ between ChR2 rats (closed) and eYFP controls (open). Repeated measures two-way ANOVA, $F_{1,11} = 1.35$, $P = 0.2705$. **r**, During competition sessions, the average time spent in the reward port for CS^{shk} trials during ON sessions and CS^{shk} trials during OFF sessions was not different between ChR2 (closed) and eYFP controls (open). Repeated measures two-way ANOVA, $F_{1,11} = 0.94$, $P = 0.354$. **s**, During competition sessions, the average time spent freezing for CS^{shk} trials during ON sessions and CS^{shk} trials during OFF sessions was not different between ChR2 rats (closed) and eYFP controls (open). Repeated measures two-way ANOVA, $F_{1,11} = 0.16$, $P = 0.6998$. Error bars and shading represent s.e.m.

Sucrose was removed from the delivery port by vacuum if not collected. **j**, VTA^{DA}-mPFC::ChR2 rats ($n = 7$) and VTA^{DA}-mPFC::eYFP controls ($n = 6$) acquired sucrose conditioning similarly. Two-way repeated measures ANOVA, $F_{2,22} = 0.7$, $P = 0.5090$. **k**, During discrimination, the alternative CS (light or tone, counterbalanced) was introduced and predicted foot shock (CS^{shk}). **l**, Average traces showing time spent in the sucrose port before, during, and after each CS presentation (grouped, $n = 13$ rats). **m**, Time spent in the sucrose port did not differ between VTA^{DA}-mPFC::ChR2 rats ($n = 7$) and VTA^{DA}-mPFC::eYFP controls ($n = 6$) during CS^{suc} or CS^{shk} presentation. Repeated measures two-way ANOVA, $F_{1,11} = 0.54$, $P = 0.4789$. **n**, Average traces showing time spent freezing before, during, and after each CS presentation (grouped, $n = 13$ rats). **o**, Time spent freezing did not differ between VTA^{DA}-mPFC::ChR2 rats and VTA^{DA}-mPFC::eYFP controls during CS^{suc} or CS^{shk} presentation. Repeated measures two-way ANOVA, $F_{1,11} = 0.01$, $P = 0.9281$. **p**, During competition sessions, the average time spent in the reward port for CS^{suc} trials during ON sessions and CS^{suc} trials during OFF sessions did not differ between ChR2 rats ($n = 7$, closed) and eYFP controls ($n = 6$, open). Repeated measures two-way ANOVA, $F_{1,11} = 0.82$, $P = 0.3845$. Note that during ON sessions, stimulation was only delivered during the CS^{comp} trials. **q**, Average time spent freezing for CS^{suc} trials during ON sessions and CS^{suc} trials during OFF sessions did not differ between ChR2 rats (closed) and eYFP controls (open). Repeated measures two-way ANOVA, $F_{1,11} = 1.35$, $P = 0.2705$. **r**, During competition sessions, the average time spent in the reward port for CS^{shk} trials during ON sessions and CS^{shk} trials during OFF sessions was not different between ChR2 (closed) and eYFP controls (open). Repeated measures two-way ANOVA, $F_{1,11} = 0.94$, $P = 0.354$. **s**, During competition sessions, the average time spent freezing for CS^{shk} trials during ON sessions and CS^{shk} trials during OFF sessions was not different between ChR2 rats (closed) and eYFP controls (open). Repeated measures two-way ANOVA, $F_{1,11} = 0.16$, $P = 0.6998$. Error bars and shading represent s.e.m.

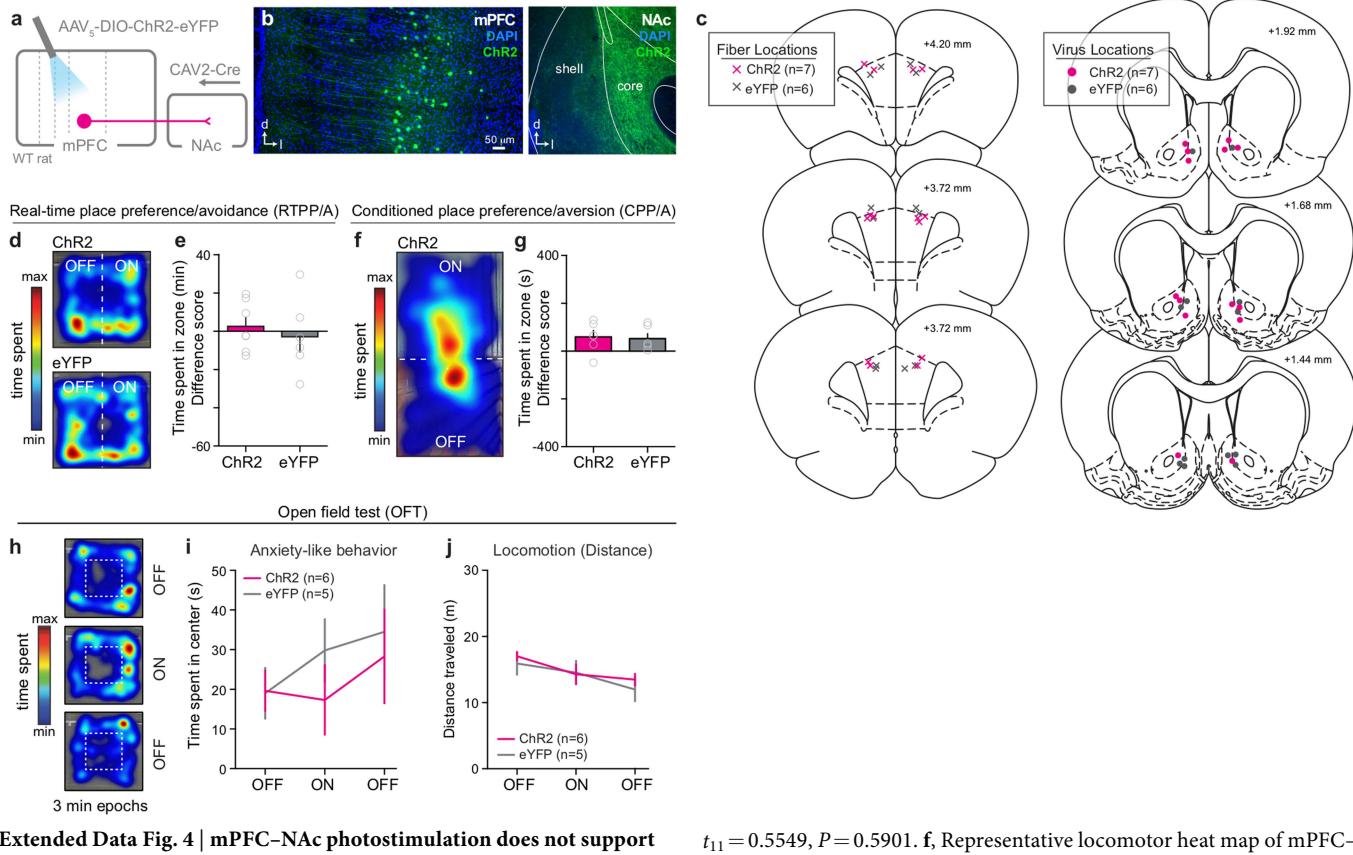


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Putative connection strength of mPFC projections to downstream targets, layer localization of projectors, and collateralization. **a**, Schematic of strategy in which anterogradely travelling virus was injected into the prelimbic and infralimbic subregions of the mPFC and fluorescence was quantified in several downstream brain regions. **b**, Orange boxes represent approximate locations of fluorescence quantification, as a proxy for connection strength. $n = 3$ rats. **c, d**, Representative images (**c**) and quantification of fluorescence (**d**) in the mPFC and downstream targets in the rat. **e**, Microinjections of CTB conjugated to fluorescent proteins (Alexa Fluor 488, Alexa Fluor 555 or Alexa Fluor 647, counterbalanced) were placed in the dPAG and NAc to retrogradely label the cell bodies of projection neurons in the rat mPFC ($n = 3$ rats). **f**, Representative confocal images of CTB injections in the NAc and dPAG of the rat. **g**, Representative confocal image of retrogradely labelled neurons in the rat mPFC. **h**, As a population, only 11 out of 1,679 CTB⁺ neurons in the mPFC were dual-labelled. **i**, Fluorescence quantification of retrogradely labelled mPFC-dPAG and mPFC-NAc neurons revealed differences in cell-body location across cortical layers in the rat mPFC. **j**, In the rat, dPAG projectors predominantly originate from deep layer 5, whereas NAc projectors are located in both superficial layers 2/3 and deep layer 5. **k**, Microinjections of CTB conjugated to fluorescent proteins were placed in the dPAG and NAc to retrogradely label the cell bodies of projection neurons in the mouse mPFC. **l**, Representative confocal images of CTB injections in the NAc and dPAG of the mouse ($n = 3$ mice). **m**, Representative confocal image of retrogradely labelled neurons in the mouse mPFC. **n**, As a population, only 17 out of 458 CTB⁺

neurons in the mPFC were dual-labelled. **o**, Fluorescence quantification of retrogradely labelled mPFC-dPAG and mPFC-NAc neurons revealed differences in cell-body location across cortical layers in the mouse mPFC. **p**, In the mouse, dPAG projectors predominantly originate from deep layer 5, whereas NAc projectors are located in both superficial layers 2/3 and deep layer 5. **q**, Schematic of viral strategy to explore downstream fluorescence from mPFC-NAc::eYFP ($n = 3$ rats) and mPFC-dPAG::eYFP ($n = 3$ rats) projectors. **r**, Quantification of fluorescence in the mPFC and downstream brain regions originating from mPFC-dPAG::eYFP and mPFC-NAc::eYFP neurons. **s, t**, Representative confocal images from a mPFC-dPAG::eYFP subject (**s**) and a mPFC-NAc::eYFP subject (**t**). **u**, Schematic of viral strategy to explore downstream terminals from mPFC-NAc::synaptophysin ($n = 3$ mice) and mPFC-dPAG::synaptophysin ($n = 3$ mice) projectors. **v**, Quantification of fluorescence in the mPFC and downstream brain regions originating from mPFC-dPAG::synaptophysin and mPFC-NAc::synaptophysin neurons. **w, x**, Representative confocal images from a mPFC-dPAG::synaptophysin subject (**w**) and a mPFC-NAc::synaptophysin subject (**x**). BLA, basolateral amygdala; agIN, agranular insula; cl, claustrum; dStr, dorsal striatum (medial); LH, lateral hypothalamus; LS, lateral septum; PHA, posterior hypothalamic area; PVT, paraventricular nucleus of the thalamus; vPAG, ventral periaqueductal grey. Rat and mouse brains in this figure have been reproduced with permission from Paxinos and Watson, 2006⁵³, and Paxinos and Franklin, 2004⁵⁴, respectively. Error bars and dashed lines represent s.e.m. Scale bars, 50 μ m.

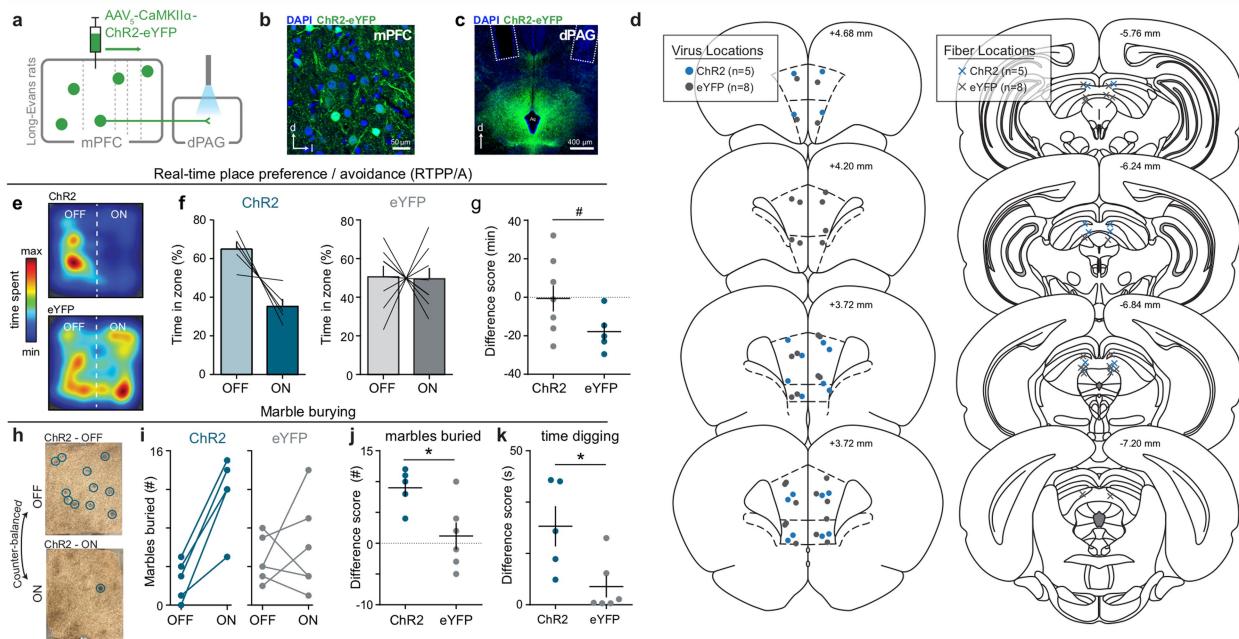
mPFC-NAc



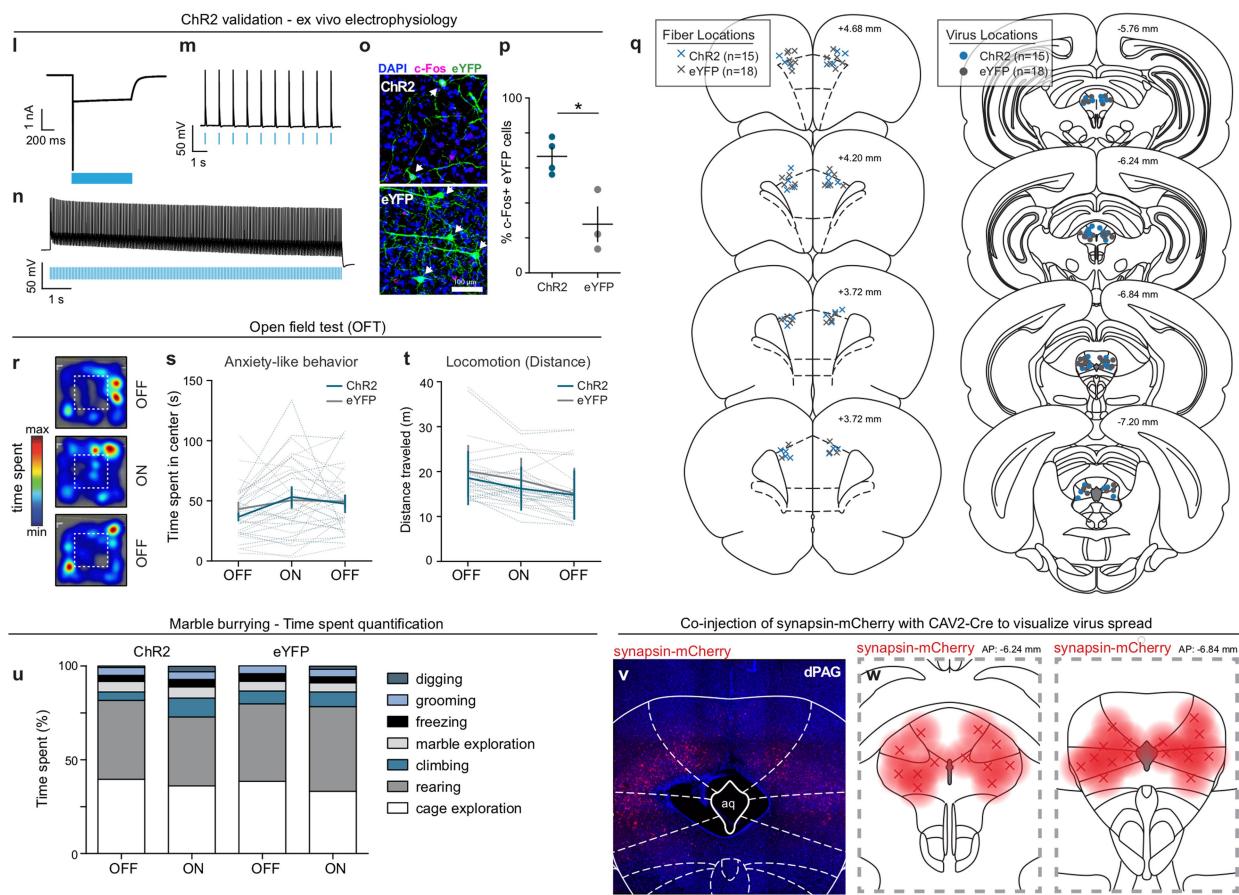
Extended Data Fig. 4 | mPFC-NAc photostimulation does not support place preference or aversion. **a**, Schematic of viral transduction strategy to achieve optogenetic control of rat mPFC neurons projecting to the NAc. **b**, Representative image of NAc-projecting mPFC neurons expressing ChR2 (left) and ChR2⁺ terminals in the NAc (right). **c**, Histological verification of bilateral optical-fibre implant locations above the mPFC and virus injection locations in the NAc. **d**, Representative locomotor heat maps of mPFC-NAc::ChR2 (top) and mPFC-NAc::eYFP (bottom) subjects in the RTPP/A assay. **e**, Optogenetic stimulation of mPFC-NAc neurons did not evoke real-time place avoidance or preference in mPFC-NAc::ChR2 animals ($n=7$ rats), compared to mPFC-NAc::eYFP controls ($n=6$ rats), measured by difference score (minutes spent in the ON zone – minutes spent in OFF zone). Two-tailed unpaired t -test, $t_{10}=0.5549$, $P=0.5901$. **f**, Representative locomotor heat map of mPFC-NAc::ChR2 subject in CPP/A assay. **g**, Optogenetic stimulation of mPFC-NAc neurons did not evoke real-time place aversion or preference in mPFC-NAc::ChR2 animals ($n=6$ rats), compared to mPFC-NAc::eYFP controls ($n=6$ rats). Two-tailed unpaired t -test, $t_{10}=0.2143$, $P=0.8346$. **h**, Representative locomotor heat maps of a mPFC-NAc::ChR2 subject during 3 min OFF-ON-OFF epochs during the open-field test. **i, j**, Optical activation of mPFC-NAc::ChR2 ($n=6$ rats) did not change time spent in the centre region compared to eYFP controls ($n=5$ rats) (**i**; two-way repeated measures ANOVA, $F_{2,18}=0.74$, $P=0.4913$), or general locomotor activity (**j**; two-way repeated measures ANOVA, $F_{2,18}=0.61$, $P=0.5532$). Data are mean \pm s.e.m. The rat brains in this figure were reproduced with permission from Paxinos and Watson, 2006⁵³.

$t_{11}=0.5549$, $P=0.5901$. **f**, Representative locomotor heat map of mPFC-NAc::ChR2 subject in CPP/A assay. **g**, Optogenetic stimulation of mPFC-NAc neurons did not evoke real-time place aversion or preference in mPFC-NAc::ChR2 animals ($n=6$ rats), compared to mPFC-NAc::eYFP controls ($n=6$ rats). Two-tailed unpaired t -test, $t_{10}=0.2143$, $P=0.8346$. **h**, Representative locomotor heat maps of a mPFC-NAc::ChR2 subject during 3 min OFF-ON-OFF epochs during the open-field test. **i, j**, Optical activation of mPFC-NAc::ChR2 ($n=6$ rats) did not change time spent in the centre region compared to eYFP controls ($n=5$ rats) (**i**; two-way repeated measures ANOVA, $F_{2,18}=0.74$, $P=0.4913$), or general locomotor activity (**j**; two-way repeated measures ANOVA, $F_{2,18}=0.61$, $P=0.5532$). Data are mean \pm s.e.m. The rat brains in this figure were reproduced with permission from Paxinos and Watson, 2006⁵³.

mPFC terminal stimulation in the dPAG



Cre-DIO enabled mPFC-dPAG cell body targeting

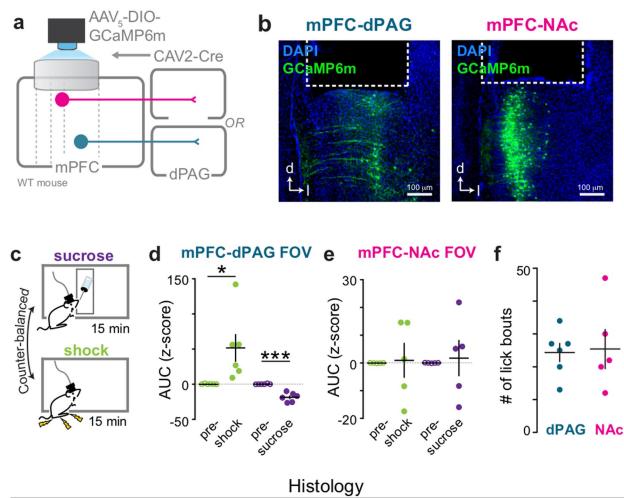


Extended Data Fig. 5 | See next page for caption.

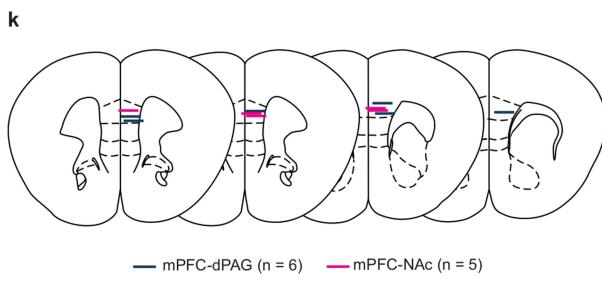
Extended Data Fig. 5 | Activation of mPFC terminals in the dPAG increases marble burying and activation of mPFC–dPAG cell bodies does not affect anxiety-like behaviour. **a**, Schematic of viral strategy to achieve optogenetic control of ChR2-expressing mPFC terminals in the dPAG. **b, c**, Representative image of ChR2⁺ neurons in the mPFC (**b**) and ChR2⁺ terminals in the dPAG (**c**) (optic fibre lesions indicated by dashed lines). **d**, Histological verification of bilateral virus injection locations in the mPFC and bilateral optic fibre implant locations above the dPAG. **e**, Representative locomotor heat maps of mPFC–dPAG::ChR2 (top) and mPFC–dPAG::eYFP control subject (bottom) in the RTPP/A assay. **f**, Percent of time spent in the ON and OFF zones of the arena in mPFC–dPAG::ChR2 and mPFC–dPAG::eYFP subjects. **g**, Optogenetic stimulation of mPFC terminals in the dPAG resulted in a trend towards avoidance in the RTPA assay in mPFC–dPAG::ChR2 animals ($n=5$ rats), compared with mPFC–dPAG::eYFP controls ($n=8$ rats). Two-tailed unpaired *t*-test, $t_{11}=1.830$, ${}^*P=0.0944$. **h**, Representative arena of mPFC–dPAG::ChR2 animal after marble-burying assay when optical stimulation was OFF (top) and ON (bottom). **i**, Number of marbles buried in mPFC–dPAG::ChR2 ($n=5$ rats) and mPFC–dPAG::eYFP ($n=6$ rats) during OFF and ON sessions. **j, k**, Optical stimulation of mPFC–dPAG neurons resulted in more marbles buried by mPFC–dPAG::ChR2 animals, compared with mPFC–dPAG::eYFP controls (**j**; two-tailed unpaired *t*-test, $t_9=2.839$, ${}^*P=0.0194$) and more time digging (**k**; one-tailed unpaired *t*-test, $t_9=2.775$, ${}^*P=0.0108$). **l**, Functional ChR2 expression in mPFC–dPAG neurons was verified by targeted ex vivo whole-cell patch-clamp electrophysiology. Recording from a ChR2-expressing mPFC–dPAG neuron in voltage-clamp mode showing sustained inward current elicited by a 1-s pulse of 470-nm light.

m, n, In current-clamp mode, action potentials were elicited by 1-Hz (**m**) and 20-Hz light trains (**n**). 470 nm, 5-ms pulse duration. **o**, Representative confocal images of mPFC–dPAG::ChR2 (top) and mPFC–dPAG::eYFP expressing neurons showing immediate early gene (c-Fos) expression following 5 min blue (473 nm) light exposure (20 Hz, 5-ms pulse duration, 15 mW). **p**, Laser light stimulation (473 nm) enhanced the number of c-Fos-positive ChR2-expressing mPFC–dPAG neurons compared with control mPFC–dPAG::eYFP neurons. mPFC–dPAG::ChR2, $n=4$ rats; mPFC–dPAG::eYFP, $n=3$ rats; two-tailed unpaired *t*-test, $t_5=3.707$, ${}^*P=0.014$. **q**, Histological verification of bilateral optical-fibre implant locations above the mPFC and virus injection locations in the dPAG for mPFC–dPAG::ChR2/eYFP-expressing rats. **r**, Representative locomotor heat maps of a mPFC–dPAG::ChR2 subject during 3 min OFF-ON-OFF epochs in the open-field test. **s, t**, Optical activation of mPFC–dPAG::ChR2 ($n=15$ rats) did not change time spent in the centre region compared to eYFP controls (**s**; $n=18$ rats, two-way repeated measures ANOVA, group \times epoch, $F_{2,62}=0.37$, $P=0.69$), or general locomotor activity (**t**; distance travelled, two-way repeated measures ANOVA, group \times epoch interaction, $F_{2,62}=0.9412$, $P=0.3957$). **u**, Quantification of behaviours (percentage of time engaging) during marble-burying assay. **v**, Representative confocal image of viral spread in the PAG, visualized by co-injection of AAV₅-hSyn-mCherry (hSyn, synapsin, red) with CAV2-Cre in a subset of mPFC–dPAG::ChR2/eYFP expressing rats. **w**, Illustration of reconstructed injection locations and spread in co-injected subjects. $n=14$ total, 7 ChR2, 7 eYFP. Error bars indicate s.e.m. The rat brains in this figure were reproduced with permission from Paxinos and Watson, 2006⁵³.

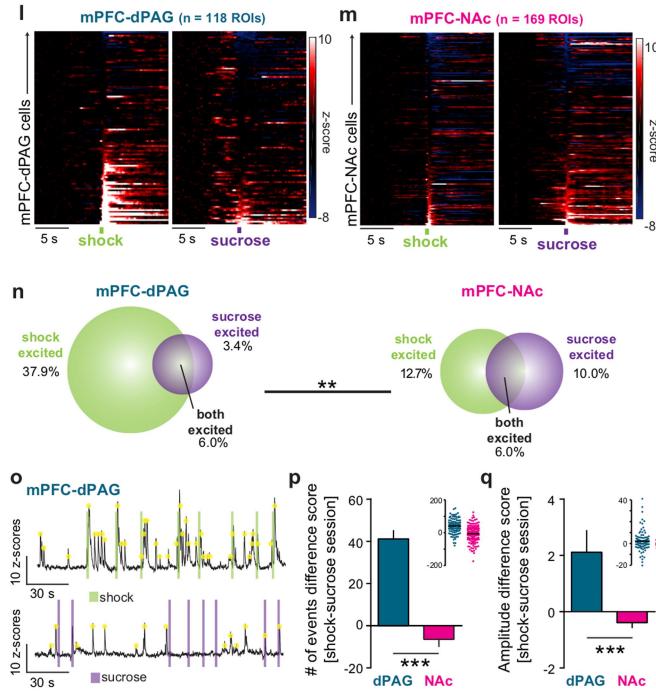
Additional FOV & behavioral results



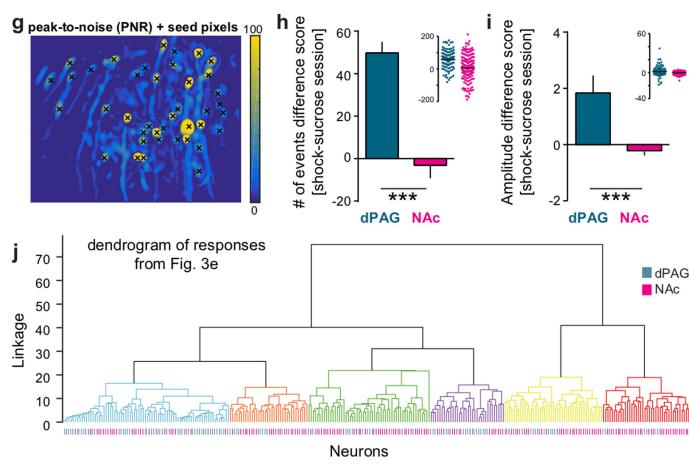
Histology



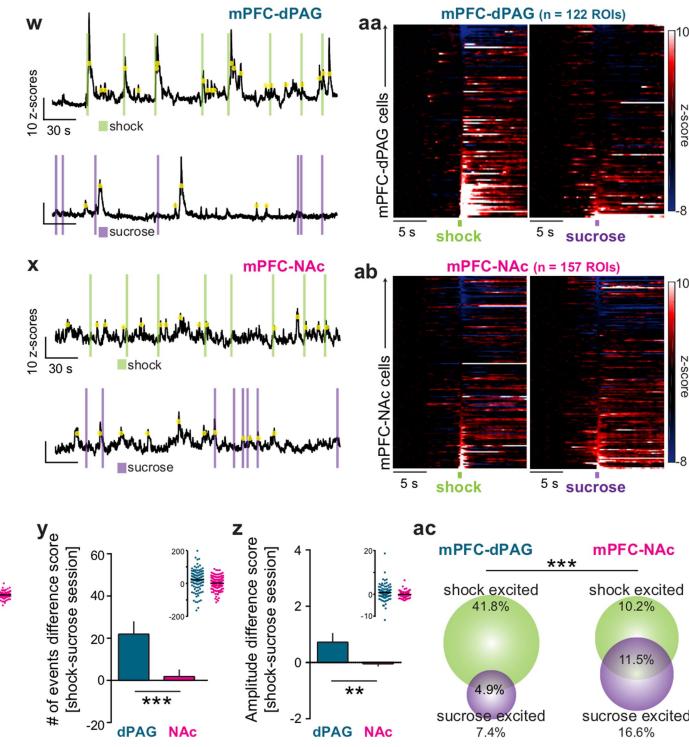
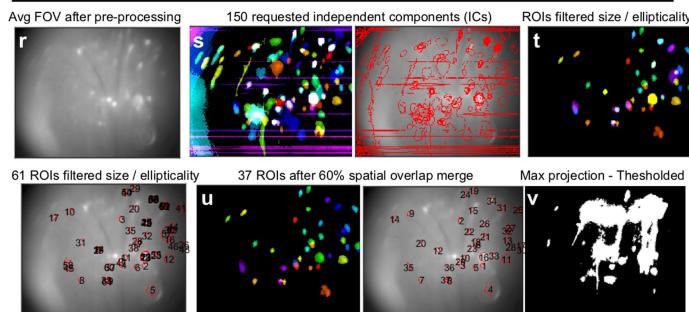
Results using CNMF-E with removal of non-negative constraints



Additional CNMF-E results



Results using "Non-ROI thresholded subtraction" method



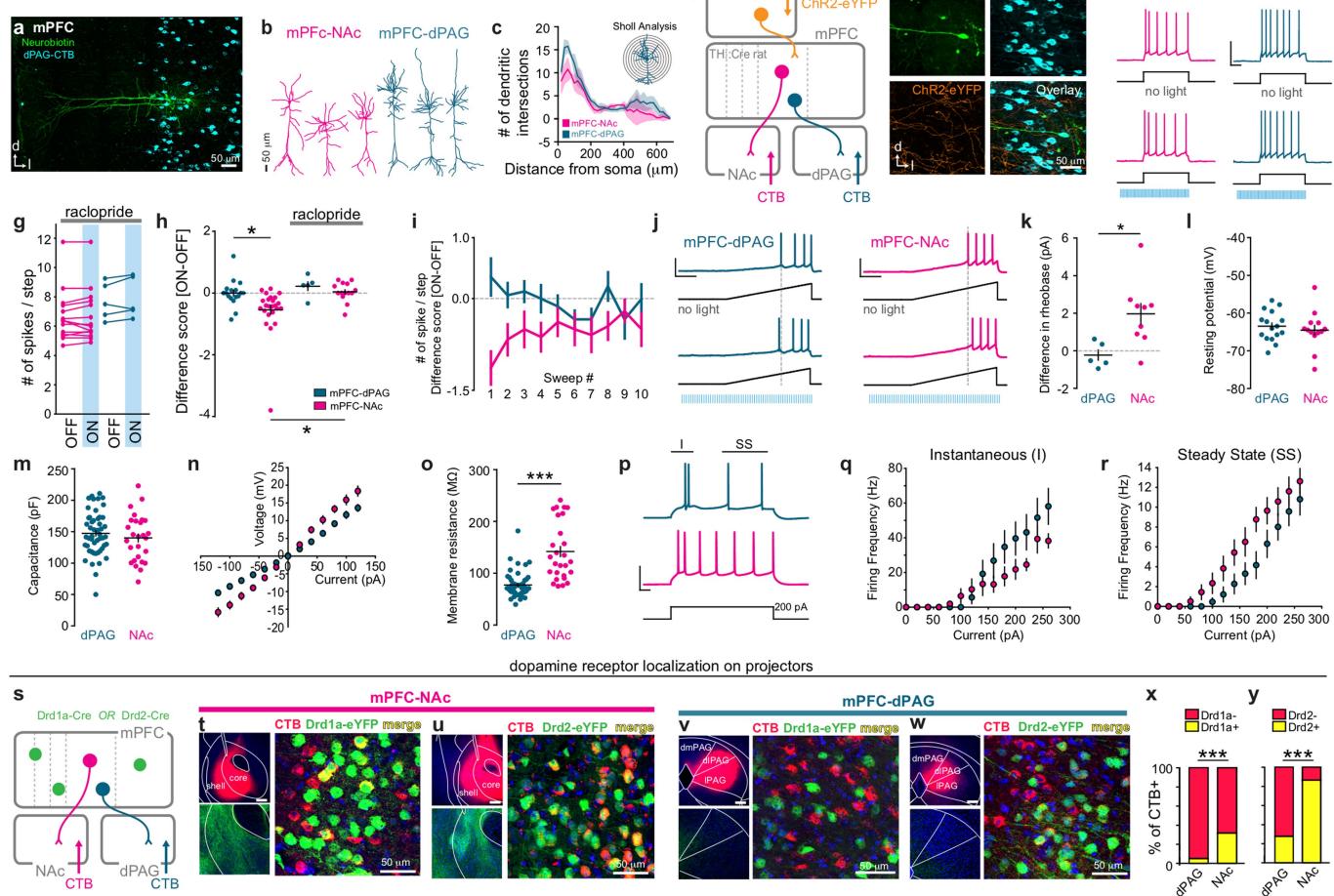
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Analysis and additional data from epifluorescent calcium imaging experiments during sucrose and shock delivery.

a, Schematic of strategy for monitoring neuronal activity in mPFC-dPAG and mPFC-NAc neurons using *in vivo* calcium imaging. **b**, Representative confocal images of mPFC-NAc::GCaMP6m (left) and mPFC-dPAG::GCaMP6m neurons (right) underneath GRIN lenses (dashed lines). **c**, Dynamic calcium fluctuations were monitored during a 15-min recording session in which mice were allowed to self-administer sucrose via a sucrose lickometer or had random, unsignalled foot shocks delivered. **d**, As a population, mPFC-dPAG::GCaMP6m ($n = 6$ mice) were activated to foot shock (green, two-tailed paired *t*-test, $t_5 = 2.616$, ${}^*P = 0.0473$) or inhibited by the initiation of a sucrose bout (purple, two-tailed paired *t*-test, $t_5 = 6.982$, ${}^{***}P = 0.0009$) as measured by the bulk fluorescence across the entire FOV (-3 to 0 s, pre-shock/sucrose; 0 – 3 s, shock/sucrose). **e**, As a population, mPFC-NAc::GCaMP6m ($n = 5$ mice) were not responsive to foot shock (green, two-tailed paired *t*-test, $t_4 = 0.1520$, $P = 0.8866$) or the initiation of a sucrose bout (purple, two-tailed paired *t*-test, $t_4 = 0.2678$, $P = 0.8021$) (-3 to 0 s, pre-shock/sucrose; 0 – 3 s, shock/sucrose). **f**, mPFC-dPAG::GCaMP6m and mPFC-NAc::GCaMP6m mice did not differ in the number of lick bouts initiated during the sucrose session. Two-tailed unpaired *t*-test, $t_9 = 0.1666$, $P = 0.8714$. **g**, Peak-to-noise heat map generated from a representative FOV with seed pixels overlaid (black X). **h**, mPFC-dPAG::GCaMP6m neurons ($n = 118$ ROIs) had more frequent calcium transients than mPFC-NAc::GCaMP6m neurons ($n = 169$ ROIs) during the shock session. Number of events difference score (shock – sucrose): dPAG Mdn, 51.5; NAc Mdn, -6 . Two-tailed Mann-Whitney test, $U = 5,840$, ${}^{***}P < 0.0001$. **i**, mPFC-dPAG::GCaMP6m neurons had higher amplitude transients than mPFC-NAc::GCaMP6m neurons during the shock session. Amplitude of events difference score (shock – sucrose): dPAG Mdn, 0.9031; NAc Mdn, -0.3549 . Mann-Whitney test, $U = 6,672$, ${}^{***}P < 0.0001$. **j**, Dendrogram of agglomerative hierarchical clustering. Different colours represent clusters based on average responses per ROI to footshock and sucrose. **k**, Histologically verified locations of GRIN lens implants. **l–ac**, In addition to using CNMF-E, imaging data were analysed using two other approaches: 1) a modified constrained CNMF-E algorithm considering calcium fluctuations can have negative transients, associated with a decrease in firing^{24,55} (for the approach, we did not constrain temporal components to >0) and 2) a ROI-based method (that is, ‘non-ROI’, **r–ac**). **l**, **m**, Calcium signals were extracted from individual ROIs and the average calcium traces per ROI were aligned to shock and sucrose bout onset for mPFC-NAc::GCaMP6m (**l**) and mPFC-dPAG::GCaMP6m recordings (**m**). **n**, The distribution of shock- and sucrose-excited cells for mPFC-dPAG::GCaMP6m neurons was different from mPFC-NAc::GCaMP6m neurons. $\chi^2 = 10.95$, ${}^{**}P = 0.0042$. **o**, Representative calcium traces from a mPFC-dPAG::GCaMP6m neuron during shock (top) and sucrose (bottom) recording sessions. Individual

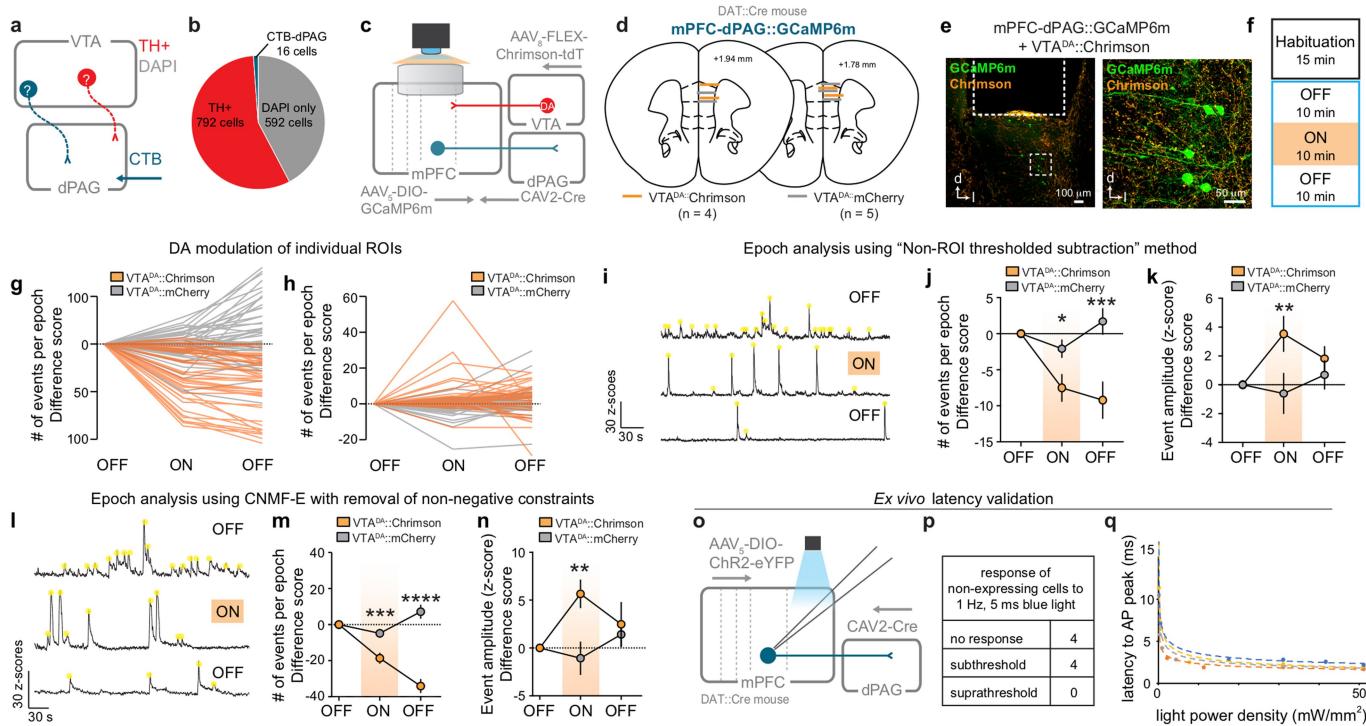
calcium transients (yellow dots) were identified and quantified. **p**, mPFC-dPAG::GCaMP6m neurons ($n = 118$ ROIs) had more frequent calcium transients than mPFC-NAc::GCaMP6m neurons ($n = 169$ ROIs) during the shock session. Number of events difference score (shock – sucrose): dPAG Mdn, 43; NAc Mdn, -3 . Two-tailed Mann-Whitney test, $U = 4,373$, ${}^{***}P < 0.0001$. **q**, mPFC-dPAG::GCaMP6m neurons had higher amplitude calcium transients compared to mPFC-NAc::GCaMP6m neurons during the shock session. Amplitude of events difference score (shock – sucrose): dPAG Mdn, 1.329; NAc Mdn, -0.2459 . Two-tailed Mann-Whitney test, $U = 7,164$, ${}^{***}P < 0.0001$. **r**, Mean *t*-projection image of the entire FOV through the relay lens after image pre-processing. Recordings were converted to changes in fluorescence compared to background fluorescence ($F - F_0$)/ F_0 using the mean *t*-projection image as reference (F_0). **s**, Calcium signals arising from ROIs were identified using independent and principal component analyses (PCA/ICA). **t**, Identified PCA/ICA filters were thresholded at their half-maximum values to define possible ROIs and were screened for neuronal morphology. ROIs were only accepted if the threshold filters included only on contiguous region with an eccentricity of <0.85 and an area between 30–350 pixels. In this example, 61 ROIs (of the original 150 independent components (ICs)) met these criteria. **u**, Accepted ROI filters were then merged if their areas overlapped by more than 60%. In this example, 24 ROIs were merged for a remaining total of 37 valid ROIs. **v**, To acquire the non-ROI thresholded image for background subtraction, max *z* projections of individual recordings were created and thresholded to separate ROIs and their processes from the rest of the FOV. Average signal from the remaining pixels was used as a proxy for the whole-field changes in fluorescence, and regressed from the signal extracted from each ROI. **w**, **x**, Calcium transients (yellow dots) within individual mPFC-dPAG::GCaMP6m neurons (**w**) and mPFC-NAc::GCaMP6m neurons (**x**) were quantified (representative traces). **y**, mPFC-dPAG::GCaMP6m neurons ($n = 113$ of 118 ROIs) had more frequent calcium transients than mPFC-NAc::GCaMP6m neurons ($n = 157$ ROIs) during the shock session. Difference score (shock – sucrose): dPAG Mdn, 30; NAc Mdn, 6. Two-tailed Mann-Whitney test, $U = 6,392$, ${}^{***}P < 0.0001$. **z**, mPFC-dPAG::GCaMP6m neurons had calcium transients of larger amplitude than mPFC-NAc::GCaMP6m neurons during the shock session. Difference score (shock – sucrose): dPAG Mdn, 0.5158; NAc Mdn, -0.0615 . Two-tailed Mann-Whitney test, $U = 7,065$, ${}^{**}P = 0.0044$. **aa**, **ab**, Average calcium traces per cell for mPFC-dPAG::GCaMP6m neurons (**aa**) and mPFC-NAc::GCaMP6m neurons (**ab**) were aligned to shock (left) and sucrose bout (right). **ac**, The distribution of shock- and sucrose-excited cells for mPFC-dPAG::GCaMP6m ($n = 118$ ROIs) neurons was different from that for mPFC-NAc::GCaMP6m neurons ($n = 157$ ROIs). $\chi^2 = 32.33$, ${}^{***}P < 0.0001$. Error bars and ‘+’ indicate s.e.m. Scale bar, 100 μ m. The mouse brains in this figure were reproduced with permission from Paxinos and Franklin, 2004⁵⁴.

Sholl analysis on reconstructed mPFC-NAc and mPFC-dPAG neurons



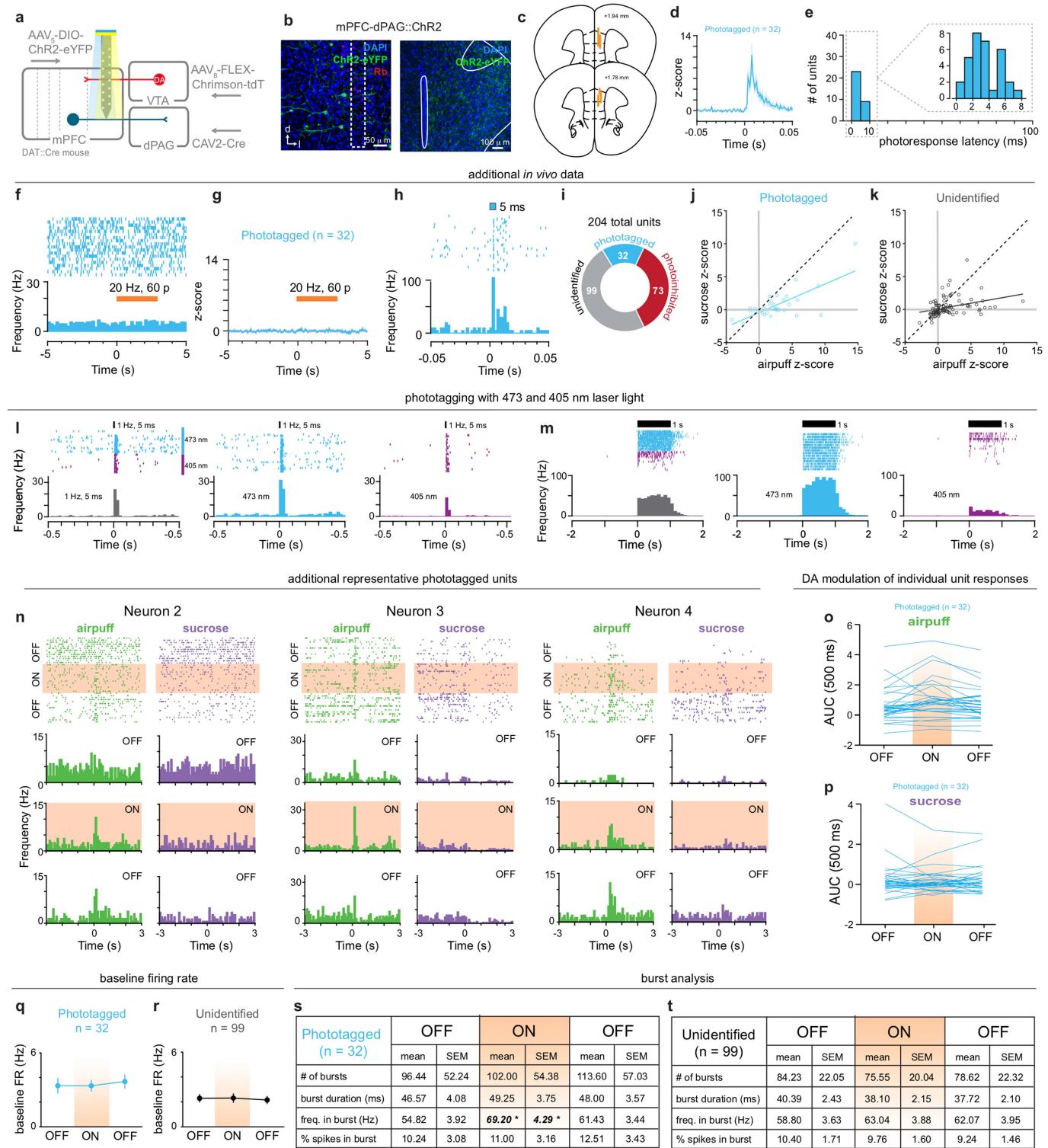
Extended Data Fig. 7 | VTA^{DA} effects on mPFC projectors across time and their properties. **a**, Representative confocal image of mPFC-dPAG labelled neurons. **b**, Representative examples of reconstructed mPFC-NAc and mPFC-dPAG neurons. **c**, Sholl analysis of mPFC-NAc ($n = 4$ cells) and mPFC-dPAG ($n = 4$ cells) subpopulations. **d**, Schematic of viral strategy to optically manipulate ChR2-expressing VTA^{DA} terminals in the mPFC and record from dPAG- and NAc-projectors retrogradely labelled with CTB using ex vivo electrophysiology. **e**, Representative images of a recorded mPFC-dPAG neuron (neurobiotin⁺ and CTB⁺) surrounded by ChR2-eYFP⁺ VTA^{DA} terminals. **f**, Representative traces of a mPFC-dPAG and mPFC-NAc neuron during a current step without (top) and with (bottom) optogenetic activation of VTA^{DA} terminals in the presence of type D2-type dopamine receptor blockade by bath-applied raclopride. **g**, The change in spike number with optical stimulation (ON-OFF) recorded from mPFC-dPAG ($n = 5$ cells) and mPFC-NAc neurons ($n = 14$ cells) in the presence of D2-receptor antagonism. **h**, The change in spike number with optical stimulation (ON-OFF) was different between mPFC-dPAG ($n = 17$ cells) and mPFC-NAc neurons ($n = 24$ cells) and was blocked by D2-receptor antagonism. One-way ANOVA, $F_{3,56} = 5.343$, $P = 0.0026$; Bonferroni multiple comparisons tests: dPAG vs NAc, $*P = 0.0040$; NAc vs NAc + raclopride, $*P = 0.0034$. **i**, Change in the number of spikes per step with optical stimulation (ON-OFF) for individual sweeps. mPFC-NAc neurons exhibited a more robust decrease in spike number during VTA^{DA} terminal stimulation during the first few sweeps, an effect that diminished in later sweeps. **j**, Representative traces showing firing elicited in mPFC-dPAG and mPFC-NAc neurons in response to current ramp with and without VTA^{DA} terminal stimulation (grey dashed line indicates time of first action potential without optical stimulation). Scale bars, 50 mV, 500 ms. **k**, Optical stimulation of VTA^{DA} terminals increased the current required to elicit an action potential (rheobase) in NAc projectors. The change in rheobase with optical stimulation (ON-OFF) was different between dPAG projectors ($n = 5$ cells) and NAc projectors ($n = 9$ cells). Two-tailed

unpaired *t*-test, $t_{12} = 2.669$, $P = 0.0205$. **l**, **m**, Neither resting membrane potential (mPFC-dPAG, $n = 16$ cells; mPFC-NAc, $n = 13$ cells) (**l**) nor capacitance (**m**) differed between dPAG-projectors ($n = 50$ cells) and NAc-projectors ($n = 27$ cells). Resting membrane potential: two-tailed unpaired *t*-test, $t_{27} = 0.6265$, $P = 0.5363$; capacitance: two-tailed unpaired *t*-test, $t_{75} = 0.8643$, $P = 0.3902$. **n**, The current-voltage (*I*-*V*) relationship of mPFC-dPAG ($n = 16$ cells) and mPFC-NAc ($n = 13$ cells) neurons obtained by applying a series of current steps in voltage-clamp mode. Two-way ANOVA, $F_{12,324} = 10.16$, $P < 0.0001$. **o**, The membrane resistance was significantly greater in NAc projectors ($n = 27$ cells) compared to dPAG projectors ($n = 50$ cells). Two-tailed unpaired *t*-test, $t_{75} = 7.030$, $***P < 0.0001$. **p**, Representative traces showing action potential firing in mPFC-dPAG and mPFC-NAc neurons in response to a depolarizing current step. Scale bars, 50 mV, 500 ms. **q**, **r**, Instantaneous (**l**) (**q**) and steady-state (SS) (**r**) firing frequency in dPAG and NAc projectors in response to increasing current steps. **s**, Schematic of strategy for identifying dopamine type 1 receptor (D1) and dopamine type 2 receptor (D2) on mPFC-projector populations using transgenic mice (Drd1a-Cre ($n = 3$ mice) and Drd2-Cre ($n = 3$ mice)), retrograde labelling, and Cre-dependent eYFP recombination. **t**, **u**, Representative confocal images of NAc CTB injections sites (upper left), mPFC terminal fluorescence (lower left), and mPFC-NAc cell bodies (right) in a Drd1a-Cre:eYFP mouse (**t**) and Drd2-Cre:eYFP mouse (**u**). **v**, **w**, Representative confocal images of dPAG CTB injections sites (upper left), mPFC terminal fluorescence (lower left), and mPFC-dPAG cell bodies (right) in a Drd1a-Cre:eYFP mouse (**v**) and a Drd2-Cre:eYFP mouse (**w**). **x**, 5% of mPFC-dPAG CTB⁺ neurons were Drd1a⁺ (19/378), whereas 31.5% of mPFC-NAc CTB⁺ neurons were co-labelled as Drd1a⁺ (151/479) (D1 $\chi^2 = 93.29$, $***P < 0.0001$). **y**, 27.6% of mPFC-dPAG CTB⁺ neurons were Drd2⁺ (74/342), whereas 86.3% of mPFC-NAc CTB⁺ neurons were co-labelled as Drd2⁺ (414/480) (D2 $\chi^2 = 345.6$, $***P < 0.0001$). Error bars, shading, and '+' represent s.e.m.



Extended Data Fig. 8 | Investigation of VTA projections to the dPAG for simultaneous epifluorescent imaging in mPFC-dPAG neurons and excitation of VTA^{DA} terminals. **a**, To verify that VTA neurons do not project to the dPAG (to allow for CAV2-Cre mediated GCaMP6m expression in dPAG neurons and simultaneous expression of the excitatory opsin Chrimson in VTA^{DA} neurons in DAT::Cre mice), VTA slices were immunostained for tyrosine hydroxylase (TH) in rats injected with the retrograde tracer CTB in the dPAG. **b**, Of 1,400 DAPI⁺ cells counted in the VTA, 792 (56%) were TH⁺, 16 (1.1%) were CTB⁺, and 0 were TH⁺ and CTB⁺. The lack of CTB⁺ cells suggests that VTA does not make a prominent projection to the dPAG. **c**, Schematic of strategy to simultaneously image fluorescent calcium activity in mPFC-dPAG::GCaMP6m neurons and activate VTA^{DA}-mPFC. **d**, Histological verification of GRIN lens locations in the mPFC in mPFC-dPAG::GCaMP6m \times VTA^{DA}::Chrimson subjects and control mPFC-dPAG::GCaMP6m \times VTA^{DA}::mCherry subjects. **e**, Representative confocal images of mPFC-dPAG::GCaMP6m and VTA^{DA}::Chrimson expression in the mPFC. **f**, Schematic of experimental design. During the ON epoch, a 590-nm LED stimulated Chrimson expressing VTA^{DA}-mPFC (20 Hz, 60 pulses of 5 ms, every 30 s). **g**, Individual ROI transient frequency analysed with CNMF-E. **h**, Individual ROI transient amplitude analysed with CNMF-E. **i**, Representative traces from a mPFC-dPAG::GCaMP6m neuron during the OFF-ON-OFF recording epochs. Calcium transients (yellow dots) for each neuron were identified and quantified. **j**, VTA^{DA}-mPFC stimulation decreased the average calcium event frequency per neuron, during both the ON and second OFF epochs. Data normalized to first OFF epoch; two-way repeated measure ANOVA, $F_{2,184} = 9.209$, $P = 0.0002$; Bonferroni multiple comparisons tests, $P < 0.05$. **k**, VTA^{DA}-mPFC stimulation increased the average calcium event amplitude per cell during the ON epoch, an effect that recovered in the second OFF epoch. Data normalized to first OFF epoch; two-way repeated measure ANOVA, $F_{2,184} = 3.756$, $P = 0.0252$; Bonferroni multiple comparisons tests: $P < 0.05$. **l-n**, Data analysed using CNMF-E with removal of non-negative temporal constraints. Chrimson: $n = 4$ mice, 44 ROIs; mCherry: $n = 5$ mice; 50 ROIs. **l**, Representative traces from a mPFC-dPAG::GCaMP6m neuron during each 10 min OFF-ON-OFF recording epoch. Calcium transients (yellow dots) for each neuron were identified and quantified. **m**, VTA^{DA}-mPFC stimulation decreased the average calcium event frequency per neuron, during both the ON and second OFF epochs. Data normalized to first OFF epoch; two-way repeated measure ANOVA, $F_{2,184} = 43.62$, $P < 0.0001$; Bonferroni multiple comparisons tests: $P < 0.05$. **n**, VTA^{DA}-mPFC stimulation increased the average calcium event amplitude per cell during the ON epoch, an effect that recovered in the second OFF epoch. Data normalized to first OFF epoch; two-way repeated measure ANOVA, $F_{2,184} = 3.50$, $P = 0.0322$; Bonferroni multiple comparisons tests: $P < 0.05$. **o**, Schematic of viral strategy to optically manipulate ChR2-expressing VTA^{DA} terminals in the mPFC and record from mPFC-dPAG::ChR2 and non-expressing neighbouring neurons with ex vivo electrophysiology. **p**, Number of non-expressing cells with different responses to 1 Hz, 5-ms blue light delivery. **q**, Latency to action-potential peak for all ChR2-expressing cells plotted against light power density. Error bars represent s.e.m. The mouse brains in this figure were reproduced with permission from Paxinos and Franklin, 2004⁵⁴.

to first OFF epoch; two-way repeated measure ANOVA, $F_{2,184} = 9.209$, $P = 0.0002$; Bonferroni multiple comparisons tests, $P < 0.05$. **k**, VTA^{DA}-mPFC stimulation increased the average calcium event amplitude per cell during the ON epoch, an effect that recovered in the second OFF epoch. Data normalized to first OFF epoch; two-way repeated measure ANOVA, $F_{2,184} = 3.756$, $P = 0.0252$; Bonferroni multiple comparisons tests: $P < 0.05$. **l-n**, Data analysed using CNMF-E with removal of non-negative temporal constraints. Chrimson: $n = 4$ mice, 44 ROIs; mCherry: $n = 5$ mice; 50 ROIs. **l**, Representative traces from a mPFC-dPAG::GCaMP6m neuron during each 10 min OFF-ON-OFF recording epoch. Calcium transients (yellow dots) for each neuron were identified and quantified. **m**, VTA^{DA}-mPFC stimulation decreased the average calcium event frequency per neuron, during both the ON and second OFF epochs. Data normalized to first OFF epoch; two-way repeated measure ANOVA, $F_{2,184} = 43.62$, $P < 0.0001$; Bonferroni multiple comparisons tests: $P < 0.05$. **n**, VTA^{DA}-mPFC stimulation increased the average calcium event amplitude per cell during the ON epoch, an effect that recovered in the second OFF epoch. Data normalized to first OFF epoch; two-way repeated measure ANOVA, $F_{2,184} = 3.50$, $P = 0.0322$; Bonferroni multiple comparisons tests: $P < 0.05$. **o**, Schematic of viral strategy to optically manipulate ChR2-expressing VTA^{DA} terminals in the mPFC and record from mPFC-dPAG::ChR2 and non-expressing neighbouring neurons with ex vivo electrophysiology. **p**, Number of non-expressing cells with different responses to 1 Hz, 5-ms blue light delivery. **q**, Latency to action-potential peak for all ChR2-expressing cells plotted against light power density. Error bars represent s.e.m. The mouse brains in this figure were reproduced with permission from Paxinos and Franklin, 2004⁵⁴.

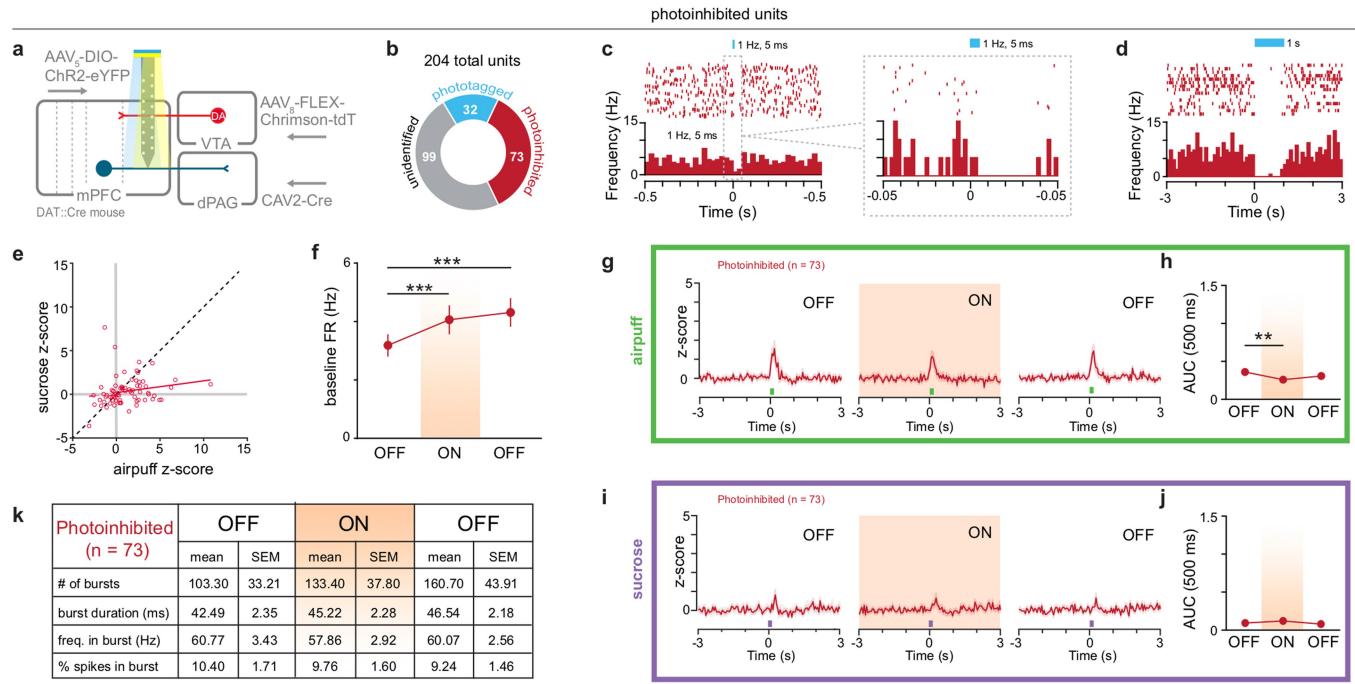


Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Additional data for head-fixed

electrophysiological recordings. **a**, Schematic of strategy to manipulate VTA^{DA} terminals in the mPFC and optically identify mPFC-dPAG::ChR2 neurons using *in vivo* head-fixed electrophysiology. **b**, Representative image of recording track in the mPFC (Rb, red retrobeads) and ChR2-eYFP-expressing mPFC-dPAG neurons. Representative image of ChR2-eYFP-expressing terminals surrounding the PAG. **c**, Histologically verified locations of recording tracks for *in vivo* head-fixed electrophysiology experiments. **d**, Population *z*-score of all phototagged units aligned to 1 Hz, 5-ms pulse of 473 nm. **e**, Photoresponse latencies showing <8 ms response latency from all 32 mPFC-dPAG::ChR2 units. **f**, **g**, PSTH from representative phototagged unit (**f**) and population *z*-score showing no response (**g**) to 20 Hz, 60 pulses of 593-nm laser light used for VTA^{DA}::Chrimson terminal activation. **h**, Representative PSTH of the firing rate in response to the onset of 5-ms pulse of 473-nm laser light used for phototagging. **i**, 204 mPFC units were recorded ($n=3$ mice, 5 recording sessions) and 32 phototagged units were identified as mPFC-dPAG projectors (blue), 73 were photoinhibited (red), and 99 remained unidentified (grey). **j**, **k**, Neural response magnitudes to airpuff (*x* axis) and sucrose (*y* axis) in phototagged (**j**; blue) and unidentified (**k**;

black) populations. **l**, **m**, In a subset of mice, both 405 and 473-nm laser light were used for phototagging. **l**, Representative phototagged unit showing faithful responses to 1 Hz, 5-ms pulses of both 473 and 405-nm light. **m**, Representative phototagged unit showing blunted response to 1 s of 405-nm, compared to 473-nm light. **n**, Representative PSTHs of phototagged mPFC-dPAG units aligned to airpuff (green) and sucrose (purple). Histograms show neural responses in the OFF-ON-OFF epochs. **o**, **p**, Individual neural responses (AUC (0–500 ms post-stimulus presentation)) of every phototagged unit ($n=32$ units) to airpuff (**o**) and sucrose (**p**) in each of the three recording epochs (OFF-ON-OFF). **q**, **r**, VTA dopamine terminal stimulation in the mPFC did not change the baseline firing rate (FR) in the 3-s pre-stimulus windows in the phototagged (**q**; Friedman test, $\chi^2=2.472$, $P=0.2905$) or unidentified (**r**; Friedman test, $\chi^2=0.4242$, $P=0.8089$) populations. **s**, VTA^{DA} terminal activation increased the frequency within a burst in the phototagged population. **t**, VTA^{DA} terminal activation did not affect burst characteristics in the unidentified population. Error bars indicate s.e.m. The mouse brains in this figure were reproduced with permission from Paxinos and Franklin, 2004⁵⁴.



Extended Data Fig. 10 | Dopamine-attenuates responses to airpuff in photoinhibited mPFC neurons. **a**, Schematic of strategy to manipulate VTA^{DA} terminals in the mPFC and optically identify mPFC-dPAG::ChR2 neurons using *in vivo* head-fixed electrophysiology. *n* = 3 mice, 5 recording sessions. **b**, 35.8% of recorded units (73/204) were photoinhibited. **c**, **d**, Representative PSTHs of a photoinhibited unit in response to 1 Hz, 5 ms (**c**) and 1 s (**d**) of 473-nm light. **e**, Neural response magnitudes to airpuff (x axis) and sucrose (y axis) in photoinhibited (red) population. **f**, VTA^{DA} terminal stimulation in the mPFC increased the baseline firing rate in the 3-s pre-stimulus windows in the photoinhibited population (*n* = 73 units) during the ON and second OFF epochs

(Friedman test, $\chi^2 = 16.22$; $P = 0.0003$; Dunn's multiple comparisons tests, $P < 0.05$). **g**, Population z-score of photoinhibited units aligned to airpuff in each of the recording epochs. **h**, In photoinhibited neurons, VTA^{DA} terminal stimulation attenuated neural responses to airpuff (Friedman test, $\chi^2 = 8.329$, $P = 0.0155$; Dunn's multiple comparisons tests, $P < 0.05$). **i**, Population z-score of photoinhibited units aligned to sucrose in each of the recording epochs. **j**, In photoinhibited neurons, VTA^{DA} terminal stimulation did not affect neural responses to sucrose (Friedman test, $\chi^2 = 0.4492$, $P = 0.7988$; Dunn's multiple comparisons tests, $P > 0.05$). **k**, VTA^{DA} terminal activation did not affect burst characteristics in the photoinhibited population. Error bars and shading represent s.e.m.

A gut microbial factor modulates locomotor behaviour in *Drosophila*

Catherine E. Schretter^{1*}, Jost Vielmetter², Imre Bartos³, Zsuzsa Marka³, Szabolcs Marka³, Sulabha Argade⁴ & Sarkis K. Mazmanian^{1*}

While research into the biology of animal behaviour has primarily focused on the central nervous system, cues from peripheral tissues and the environment have been implicated in brain development and function¹. There is emerging evidence that bidirectional communication between the gut and the brain affects behaviours including anxiety, cognition, nociception and social interaction^{1–9}. Coordinated locomotor behaviour is critical for the survival and propagation of animals, and is regulated by internal and external sensory inputs^{10,11}. However, little is known about how the gut microbiome influences host locomotion, or the molecular and cellular mechanisms involved. Here we report that germ-free status or antibiotic treatment results in hyperactive locomotor behaviour in the fruit fly *Drosophila melanogaster*. Increased walking speed and daily activity in the absence of a gut microbiome are rescued by mono-colonization with specific bacteria, including the fly commensal *Lactobacillus brevis*. The bacterial enzyme xylose isomerase from *L. brevis* recapitulates the locomotor effects of microbial colonization by modulating sugar metabolism in flies. Notably, thermogenetic activation of octopaminergic neurons or exogenous administration of octopamine, the invertebrate counterpart of noradrenaline, abrogates the effects of xylose isomerase on *Drosophila* locomotion. These findings reveal a previously unappreciated role for the gut microbiome in modulating locomotion, and identify octopaminergic neurons as mediators of peripheral microbial cues that regulate motor behaviour in animals.

Coordinated locomotion is required for fundamental activities of life such as foraging, social interaction and mating, and involves the integration of multiple contextual factors including the internal state of the animal and external sensory stimuli^{10,11}. The intestine represents a major conduit for exposure to environmental signals that influence host physiology, and is connected to the brain through both neuronal and humoral pathways. Recent studies have shown that the intestinal microbiome regulates developmental and functional features of the nervous system^{1,2}, although the effects of gut bacteria on the neuromodulators and neuronal circuits involved in locomotion remain poorly understood. As central mechanisms of locomotion, including sensory feedback and neuronal circuits that integrate these modalities, are shared in lineages spanning arthropods and vertebrates^{11–13}, we used the fruit fly *D. melanogaster* to explore host–microbiome interactions that contribute to locomotor behaviour. Locomotion was examined in the presence (conventional; Conv) and absence (axenic; Ax) of commensal bacteria. In comparison to conventionally reared animals, axenic female adult flies showed increased walking speed and daily activity (Fig. 1a, b, g). *Drosophila* locomotion is characterized by a pattern of intermittent periods of pauses and activity bouts^{11,14}, during the latter of which the average speed of the fly is above a set threshold of 0.25 mm s^{-1} . An increase in average speed may be related to changes in temporal patterns, including the number and/or duration of walking bouts¹⁴. In axenic flies, the average length of walking bouts was higher and the average pause length was lower than in conventionally reared

flies, whereas the number of bouts was the same for the two groups (Fig. 1c–f). These data reveal that the microbiota modulates walking speed and temporal patterns of locomotion in *Drosophila*.

The microbial community of *D. melanogaster* contains 5–20 bacterial species^{15,16}. In laboratory-raised flies, two of the dominant species are *L. brevis* and *Lactobacillus plantarum*¹⁵. Specific bacteria in this community affect distinct features of *Drosophila* physiology, and even closely related microbial taxa can have different biological influences on the host^{15,17,18}. Accordingly, we examined whether locomotor performance was affected differently by individual bacterial species. Despite similar levels of colonization (Extended Data Fig. 1a), mono-association with *L. brevis*—but not *L. plantarum*—starting at eclosion was sufficient to correct changes in speed and daily activity in axenic flies (Fig. 1a, b, g and Extended Data Fig. 1b–e). Varying the strain of *L. brevis* or the host diet did not alter bacterial influences on host speed (Extended Data Fig. 1c–e), and *L. brevis* could largely restore temporal patterns of locomotion (Fig. 1c–f and Extended Data Fig. 1f). Detailed gait analysis revealed that flies associated with *L. brevis* showed comparable locomotor coordination to that of conventionally reared flies (Fig. 1h and Extended Data Fig. 1g). Furthermore, axenic flies that were colonized with a 1:1 mixture of *L. brevis* and *L. plantarum* showed similar changes in walking speed to flies associated with *L. brevis* alone (Extended Data Fig. 1h).

To investigate whether the effects of microbial exposure depend on host developmental stage, we mono-colonized flies at 3–5 days post-eclosion (Extended Data Fig. 2a), when the development of the gastrointestinal tract and remodelling of the nervous system are complete^{19–21}. Colonization with *L. brevis* alone in fully developed animals decreased locomotor speed and average walking bout length to levels similar to those of flies treated immediately after eclosion (Extended Data Fig. 2b–e). Changes in locomotion are likely to be independent of bacterial effects on host development, as conventionally reared flies treated after eclosion with broad spectrum antibiotics exhibited similar walking speeds to animals born under axenic conditions (Extended Data Fig. 2f). Administration of antibiotics increased fly locomotion in two wild-type lines (Extended Data Fig. 2g). Furthermore, colonization with *L. brevis*, but not *L. plantarum*, after the removal of antibiotics reduced locomotor behaviour to levels similar to conventionally reared flies (Fig. 1i and Extended Data Fig. 2h–l). From these data, we conclude that locomotion is modulated by select bacterial species of the *Drosophila* microbiome, and is mediated by active signalling rather than developmental influences.

Gut bacteria secrete molecular products that regulate aspects of host physiology, including immunity and feeding behaviour^{22,23}. To explore how microbes influence locomotion, we administered either cell-free supernatant (CFS) collected from bacterial cultures or heat-killed bacteria to axenic flies. CFS alone from *L. brevis* (*L. b* CFS) significantly reduced hyperactivity in axenic flies whereas heat-killed bacteria did not (Fig. 2a and Extended Data Fig. 3a–e), demonstrating a requirement for metabolically active *L. brevis* for modulation of locomotion.

¹Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. ²Protein Expression Center, Beckman Institute, California Institute of Technology, Pasadena, CA, USA. ³Department of Physics, Columbia University, New York, NY, USA. ⁴GlycoAnalytics Core, University of California, San Diego, CA, USA. *e-mail: cschrett@caltech.edu; sarkis@caltech.edu

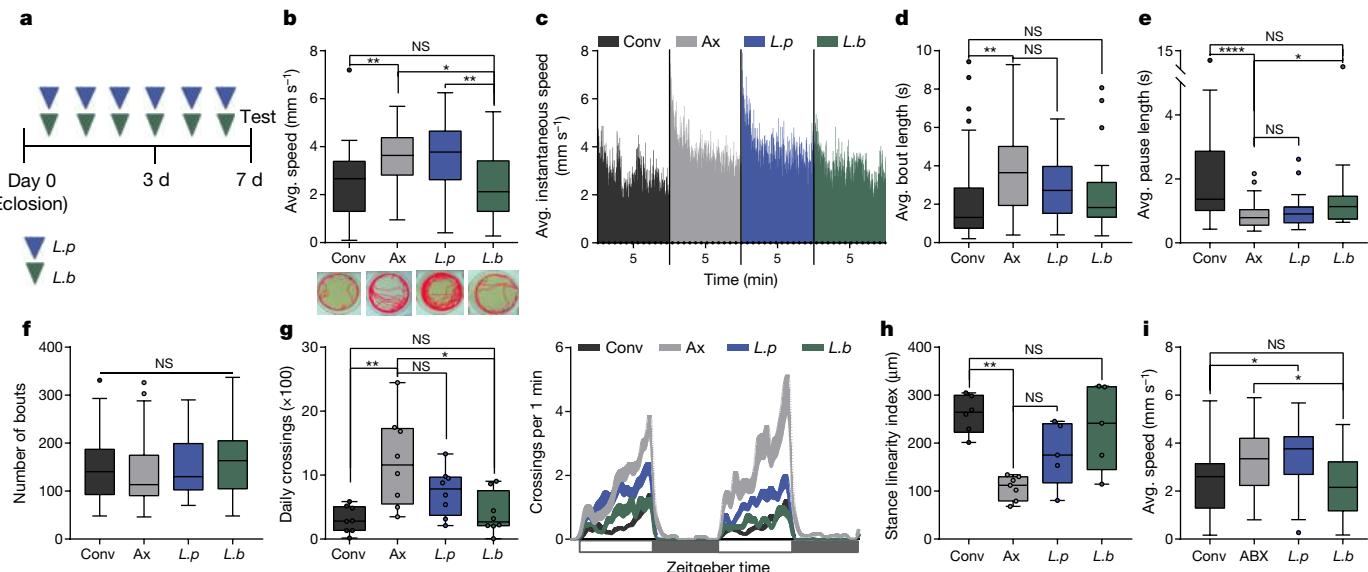


Fig. 1 | Select gut bacteria modulate locomotor behaviour in flies.

a, Experimental design unless otherwise stated. In brief, female Oregon^R flies were either left untreated or treated with a bacterial culture or bacterial-derived factors through application to the fly medium (40 μ l) daily for 6 days. **b–f**, Analysis of average speed (**b**; below are representative individual traces), average instantaneous speed (**c**; dashes show 5-min mark for each group), average bout length (**d**), average pause length (**e**) and number of bouts (**f**) of locomotion in conventional (Conv), axenic (Ax), and *L. plantarum* (*L.p*) or *L. brevis* (*L.b*) mono-associated flies over a 10-min period. **b**, Conv, n = 36 flies; Ax, n = 36; *L.p*, n = 35; *L.b*, n = 36. **c**, Conv, n = 23; Ax, n = 29; *L.p*, n = 23; *L.b*, n = 21. **d–f**, Conv, n = 32; Ax, n = 23; Ax, n = 29; *L.p*, n = 23; *L.b*, n = 21. **d–f**, Conv, n = 32; Ax, n = 23.

g, n = 36; *L.p*, n = 22; *L.b*, n = 20. **g**, Daily activity of virgin female Oregon^R flies over a 2-day light–dark (12:12 h) cycle period, starting at time 0. n = 8 flies per condition. **h**, Stance linearity index for each group. Conv, n = 6; Ax, n = 7; *L.p*, n = 5; *L.b*, n = 5. **i**, Average speed of untreated conventional flies and conventional flies treated with antibiotics (ABX) for 3 days, following which flies were either left naive or colonized with *L.p* or *L.b*. Conv, n = 25; ABX, n = 29; *L.p*, n = 24; *L.b*, n = 35. Box-and-whisker plots show median and interquartile range (IQR); whiskers show either 1.5 \times IQR of the lower and upper quartiles or range. * P < 0.05, ** P < 0.01, *** P < 0.0001. For specific P values, see Supplementary Information. Kruskal–Wallis and Dunn's post hoc tests were used for statistical analysis.

L. brevis produces uracil¹⁸, a molecule that affects the host immune response and may affect locomotion²². However, administration of physiological levels of uracil to axenic flies did not alter walking speed (Extended Data Fig. 3f). We next investigated whether immunity or feeding behaviour altered microbial-mediated locomotion. Depletion of the microbiome in immune-deficient (IMD) and Toll-knockout flies using antibiotics resulted in similar increases in walking speed compared to wild-type flies treated with antibiotics (Extended Data Fig. 4a, b). There were no differences in the expression of antimicrobial peptides or the dual oxidase gene *Duox* in axenic flies treated with *L.b* CFS compared with untreated axenic flies (Extended Data Fig. 4c). Moreover, although food intake may be influenced by bacterial species and can inhibit locomotor behaviour^{23–25}, there was no significant change in the amount of food ingested by flies treated with *L.b* CFS compared to untreated controls (Extended Data Fig. 4d, e).

Bacterial metabolism of amino acids and carbohydrates is associated with changes in host behaviour^{6,8}; however, it is not known whether bacterial metabolic enzymes influence host locomotion. We used biochemical analysis of *L.b* CFS and comparative functional analysis of bacterial strains^{26–28}, and determined that bacterial locomotor effects are mediated via proteinaceous molecule(s) present in specific bacteria, including *L. brevis* and *Escherichia coli* (Extended Data Fig. 5a–e). Subsequently, a screen of *E. coli* strains containing single-gene mutations related to amino acid and carbohydrate metabolism identified xylose isomerase (Xi) as a candidate factor for modulation of locomotor behaviour (Extended Data Fig. 5f). Xi catalyses the reversible isomerization of certain sugars, including the conversion of D-glucose to D-fructose²⁹, and is present only in *L. brevis* and *E. coli* out of the bacterial strains tested (Extended Data Fig. 5e). Administration of His-tagged Xi from *L. brevis* (Xi*) reduced locomotor behaviour in axenic flies to levels similar to those in axenic flies treated with *L.b* CFS and conventionally reared flies (Fig. 2b, c and Extended Data Fig. 5g, h). The addition of His-tagged L-arabinose isomerase (Ai*), an enzyme that is not differentially expressed among the bacteria tested, did not influence

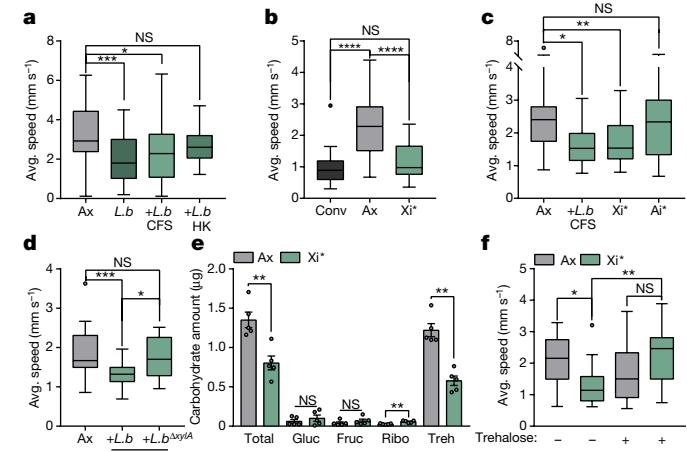


Fig. 2 | Xylose isomerase from *L. brevis* alters host locomotion.

a–c, Average speed of conventional, axenic, *L.b* mono-associated, *L.b* CFS-treated axenic, heat-killed (HK) *L.b*-treated axenic, Xi* (100 μ g ml⁻¹)-treated axenic and Ai* (100 μ g ml⁻¹)-treated axenic flies. **a**, Ax, n = 57 flies; *L.b*, n = 36; *L.b* HK, n = 24; *L.b* CFS, n = 42; Xi*, n = 29. **b**, Conv, n = 17; Ax, n = 45; Xi*, n = 29. **c**, Ax, n = 31; *L.b* CFS, n = 12; Xi*, n = 28; Ai*, n = 13. **d**, Average speed of axenic flies and axenic flies treated with CFS from wild-type *L.b* or *xybA* mutant *L.b* (*L.b*^{AxybA}) strains. Ax, n = 28; *L.b* CFS, n = 29; *L.b*^{AxybA} CFS, n = 18. **e**, Carbohydrate levels (mean \pm s.e.m.) in axenic and Xi*-treated flies. n = 5 samples per condition. **f**, Average speed of axenic flies and Xi*-treated axenic flies left untreated or supplemented with trehalose (10 mg ml⁻¹) for 3 days before testing. Ax, n = 16; Xi*, n = 18; Ax + Treh, n = 16; Xi* + Treh, n = 17. Box-and-whisker plots show median and IQR; whiskers show 1.5 \times IQR of the lower and upper quartiles. * P < 0.05, ** P < 0.01, *** P < 0.001, **** P < 0.0001. For specific P values, see Supplementary Information. Kruskal–Wallis and Dunn's (**a–d**, **f**) or Mann–Whitney *U* (**e**) post hoc tests were used for statistical analysis. Gluc, glucose; Fruc, fructose; Ribo, ribose; Treh, trehalose.

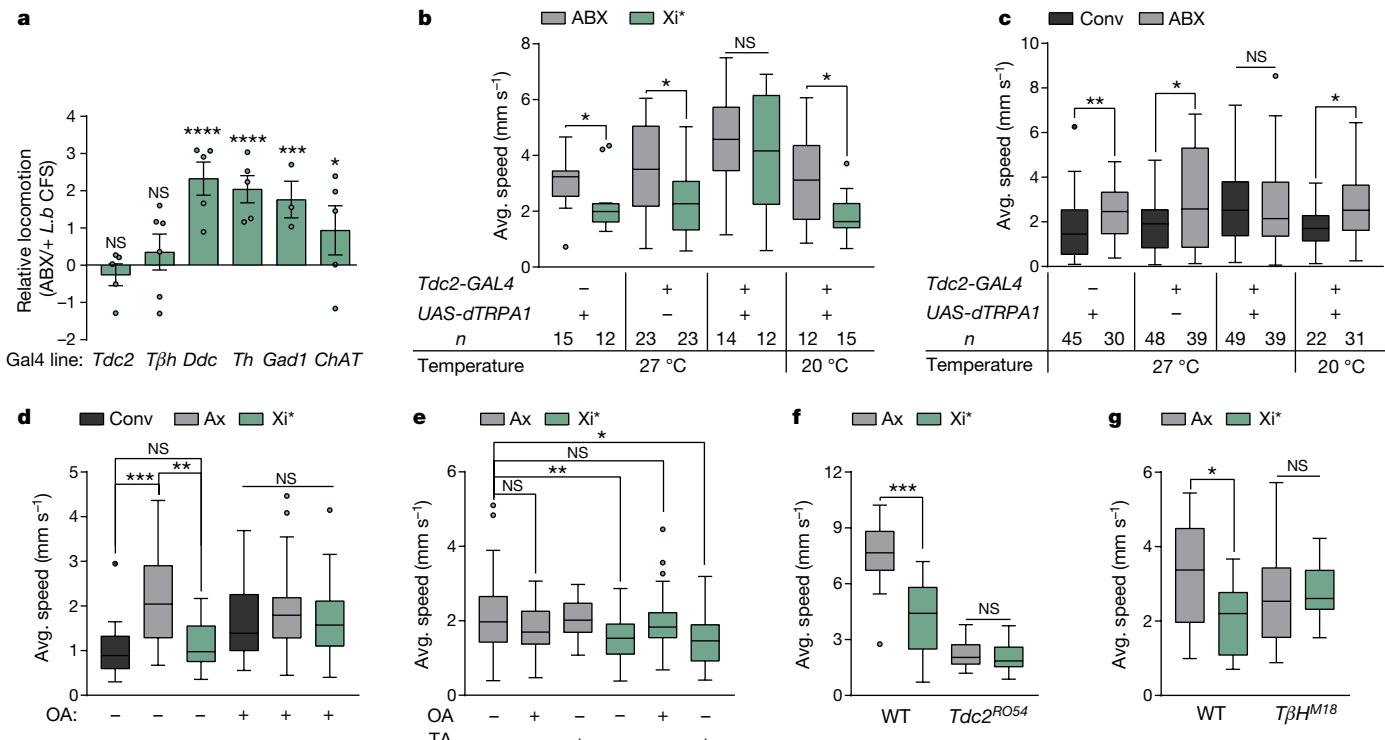


Fig. 3 | Octopamine mediates Xi-induced changes in locomotion.

a, Difference in average speed (mean \pm s.e.m.) for each GAL4 line crossed with *UAS-dTRPA1* at 27 °C. Each point denotes an independent trial. *Tdc2*, *Ddc*, *Th* (also known as *Ple*), *ChAT*, $n = 5$ trials; *Tβh*, $n = 6$; *Gad1*, $n = 3$. **b, c**, Average speed with or without thermogenetic activation. **d, e**, Average speed of flies left untreated or supplemented with octopamine (OA) or tyramine (TA) daily for 3 days. **d**, Conv, $n = 13$; Ax, $n = 33$; Xi*, $n = 21$; Conv + OA, $n = 29$; Ax + OA, $n = 27$; Xi* + OA, $n = 32$. **e**, Ax, $n = 58$; Ax + OA, $n = 13$; Ax + TA, $n = 10$; Xi*, $n = 54$; Xi* + OA, $n = 46$; Xi* + TA, $n = 27$. **f, g**, Average speed of antibiotic-treated wild-type (WT), *Tdc2*-null mutants (*Tdc2*^{RO54}) or *Tβh*-null

mutants (*Tβh*^{M18}) left untreated (ABX) or treated with Xi* daily for 3 days. **f**, WT (w+): ABX, $n = 12$; Xi*, $n = 17$; *Tdc2*^{RO54}: ABX, $n = 19$; Xi*, $n = 17$. **g**, WT (Canton-S): $n = 15$; *Tβh*^{M18}: ABX, $n = 11$; Xi*, $n = 12$. Box-and-whisker plots show median and IQR; whiskers show 1.5 \times IQR of the lower and upper quartiles. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$. For specific P values, see Supplementary Information. Mann–Whitney U post hoc tests following a two-way ANOVA (**a–c**), Kruskal–Wallis and Dunn's post hoc tests (**d, e**), or Mann–Whitney U post hoc tests (**f, g**) were used for statistical analysis. *Ddc*, DOPA decarboxylase; *Th*, tyrosine hydroxylase; *Gad1*, glutamate decarboxylase 1; *ChAT*, choline acetyltransferase.

speed in axenic flies (Fig. 2c). Furthermore, we generated a chromosomal deletion of the xylose isomerase gene *xylA* in *L. brevis*, and found that the mutant strain lacked the ability to modulate host speed and daily activity (Fig. 2d and Extended Data Fig. 5g). No changes in survival or intestinal cellular apoptosis occurred at the time of motor testing (Extended Data Fig. 5i, j). In addition, treatment with Xi* did not significantly alter sleep in axenic flies (Extended Data Fig. 6). Neither the addition of predicted products of Xi (D-fructose, D-glucose, D-xylose and D-xylulose) alone nor treatment with Xi inactivated by EDTA²⁹ reduced walking speed in axenic flies (Extended Data Fig. 7a–c). Carbohydrate analysis of whole flies showed that flies given Xi* had increased levels of ribose and reduced levels of trehalose compared to axenic controls (Fig. 2e), with no differences in these sugars in the fly medium (Extended Data Fig. 7d). Whereas EDTA-treated Xi* did not significantly alter trehalose levels, flies treated with this still displayed heightened levels of ribose compared to axenic controls (Extended Data Fig. 7e). In addition, similar to previous findings³⁰, conventionally reared and *L. brevis*-colonized flies showed reduced levels of trehalose compared to axenic flies (Extended Data Figs. 7f, g). Administration of trehalose alone reversed the effects of the microbiota on host speed, whereas supplementation with arabinose or ribose did not (Fig. 2f and Extended Data Fig. 7h–k). Collectively, these results demonstrate that Xi from *L. brevis* is sufficient to control locomotion in *Drosophila*, probably via modulation of key carbohydrates such as trehalose.

Specific neuronal pathways regulate complex behaviours in animals^{31–33}, and can be modulated by peripheral inputs such as intestinal and circulating carbohydrate levels³⁴. To explore the involvement of various neuronal subsets in bacteria-induced motor behaviour, we

used the thermosensitive cation channel *Drosophila* TRPA1 (dTRPA1) to activate neuronal populations that have been implicated in locomotion³⁵, via a repertoire of GAL4-driver lines. In combination with *UAS-dTrpA1* at the activity-inducing temperature (27 °C), activation of only two GAL4 lines—tyrosine decarboxylase (*Tdc2*) and tyramine beta-hydroxylase (*Tβh*), both of which label octopaminergic neurons—overrode modulation of locomotion by *L. brevis* (Fig. 3a and Extended Data Fig. 8). Accordingly, activation of *Tdc*-expressing cells abrogated the effects of Xi* treatment and removed the difference between conventionally reared and antibiotic-treated flies (Fig. 3b, c and Extended Data Fig. 9). The ability of *L. brevis* to decrease locomotion, however, was not changed by the activation of dopaminergic, serotoninergic, γ -aminobutyric acid (GABA)ergic or cholinergic neurons (Fig. 3a and Extended Data Fig. 8e–h). The administration of octopamine to conventionally reared, Xi*-, or *L. b* CFS-treated flies increased host walking speed to levels similar to that of axenic flies (Fig. 3d, e and Extended Data Fig. 10a). Furthermore, levels of *Tdc2* and *Tβh* (also known as *Tbh*) transcripts were lower in RNA extracted from the heads of Xi*- and *L. b* CFS-treated flies than in RNA from axenic flies (Extended Data Fig. 10b, c). As *Tdc* and *Tβh* are important for octopamine synthesis, these results further link octopamine to Xi-induced locomotor effects. Octopamine and tyramine are involved in multiple aspects of host physiology, including metabolism and behaviour, and have opposing roles in regulating certain motor behaviours^{36–44}. Whereas administration of tyramine did not influence walking speed in axenic flies treated with Xi* or *L. b* CFS (Fig. 3e and Extended Data Fig. 10d), antibiotic-treated flies carrying a null allele for *Tdc* (*Tdc2*^{RO54}) no longer displayed differences in locomotion upon supplementation with Xi*

(Fig. 3f), suggesting that tyramine has an indirect role in this process. Limiting the expression of a transgene for diphtheria toxin (*DTI*) to octopaminergic and tyraminergic neurons outside the ventral nerve cord^{39,45} resulted in equivalent average speeds between antibiotic and *Xi*^{*}-treated flies (Extended Data Fig. 10e), implicating neurons in the supra-oesophageal and the sub-oesophageal zones in microbial effects on motor behaviour. Octopamine signalling is necessary for locomotor changes, as axenic flies treated with mianserin—an octopamine receptor antagonist—and antibiotic-treated flies carrying a null allele for *Tβh* (*TβH^{M18}*) or expressing *Tβh* RNAi no longer responded to treatment with *Xi*^{*} or *L.b* CFS (Fig. 3g and Extended Data Fig. 10f–h). Similar results were also found when conventionally reared flies were compared to antibiotic-treated groups (Extended Data Fig. 10i–k). We conclude that defined products of the microbiome, and specifically *Xi*, negatively regulate octopaminergic pathways to modulate *Drosophila* locomotion (Extended Data Fig. 10l).

The microbiome influences neurodevelopment, regulates behaviour and contributes to various neurological and neuropsychiatric disorders. We have shown that gut bacteria modulate locomotion in female *Drosophila*. Depletion of the gut microbiota increases host exploratory behaviour, and the commensal bacterium *L. brevis* is sufficient to regulate locomotion. In addition, we have established that *Xi* from *L. brevis* corrects the locomotor phenotypes of axenic flies, a process that is mediated by trehalose and octopamine signalling in the host. However, further work is needed to identify the exact neurons and neuronal mechanisms involved, including potential changes in firing patterns. It will also be important to clarify sex-specific aspects of these microbial effects on locomotion^{30,46}. Notably, germ-free mice show hyperactivity similar to that of axenic *Drosophila*, and specific bacteria have been shown to decrease locomotor activity in mice^{1,47,48}, although the neuronal pathways implicated in mammalian systems have yet to be identified. The mammalian counterpart of octopamine, noradrenaline, modulates locomotion^{31,49}, potentially implicating adrenergic circuitry as a conserved pathway that is co-opted by the microbiome in flies and mammals. In addition to motor behaviour, octopamine signalling has been linked to sugar metabolism, and trehalose serves as a major energy source for *Drosophila*³⁶. Xylose isomerase may therefore facilitate adrenergic regulation of host physiology by orchestrating metabolic homeostasis, perhaps by altering internal energy storage, although additional work is needed to define how the microbiome mediates interactions between sugar metabolism and octopamine signalling. The inextricable link between metabolic state and locomotion suggests that peripheral influences on metabolism may signal via neuronal pathways to modulate physical activity. As animals have become metabolically intertwined with their microbiomes, perhaps it is not surprising that a fundamental trait such as locomotion is influenced by host–microbe symbiosis.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0634-9>.

Received: 3 July 2017; Accepted: 11 September 2018;

Published online 24 October 2018.

- Diaz Heijtz, R. et al. Normal gut microbiota modulates brain development and behavior. *Proc. Natl Acad. Sci. USA* **108**, 3047–3052 (2011).
- Bravo, J. A. et al. Ingestion of *Lactobacillus* strain regulates emotional behavior and central GABA receptor expression in a mouse via the vagus nerve. *Proc. Natl Acad. Sci. USA* **108**, 16050–16055 (2011).
- Luczynski, P. et al. Microbiota regulates visceral pain in the mouse. *eLife* **6**, e25887 (2017).
- Gacias, M. et al. Microbiota-driven transcriptional changes in prefrontal cortex override genetic differences in social behavior. *eLife* **5**, e13442 (2016).
- Fischer, C. N. et al. Metabolite exchange between microbiome members produces compounds that influence *Drosophila* behavior. *eLife* **6**, 1–25 (2017).
- Leitão-Gonçalves, R. et al. Commensal bacteria and essential amino acids control food choice behavior and reproduction. *PLoS Biol.* **15**, e2000862 (2017).
- Wong, A. C. N. et al. Gut microbiota modifies olfactory-guided microbial preferences and foraging decisions in *Drosophila*. *Curr. Biol.* **27**, 2397–2404.e4 (2017).
- Liu, W. et al. Enterococci mediate the oviposition preference of *Drosophila melanogaster* through sucrose catabolism. *Sci. Rep.* **7**, 13420 (2017).
- Sharon, G. et al. Commensal bacteria play a role in mating preference of *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **107**, 20051–20056 (2010).
- Huston, S. J. & Jayaraman, V. Studying sensorimotor integration in insects. *Curr. Opin. Neurobiol.* **21**, 527–534 (2011).
- Dickinson, M. H. et al. How animals move: an integrative view. *Science* **288**, 100–106 (2000).
- Pearson, K. G. Common principles of motor control in vertebrates and invertebrates. *Annu. Rev. Neurosci.* **16**, 265–297 (1993).
- Strausfeld, N. J. & Hirth, F. Deep homology of arthropod central complex and vertebrate basal ganglia. *Science* **340**, 157–161 (2013).
- Martin, J. R., Ernst, R. & Heisenberg, M. Temporal pattern of locomotor activity in *Drosophila melanogaster*. *J. Comp. Physiol.* **184**, 73–84 (1999).
- Erkoser, B., Storelli, G., Defaye, A. & Leulier, F. Host-intestinal microbiota mutualism: “learning on the fly”. *Cell Host Microbe* **13**, 8–14 (2013).
- Wong, C. N., Ng, P. & Douglas, A. E. Low-diversity bacterial community in the gut of the fruitfly *Drosophila melanogaster*. *Environ. Microbiol.* **13**, 1889–1900 (2011).
- Schwarzer, M. et al. *Lactobacillus plantarum* strain maintains growth of infant mice during chronic undernutrition. *Science* **351**, 854–857 (2016).
- Lee, K.-A. et al. Bacterial-derived uracil as a modulator of mucosal immunity and gut-microbe homeostasis in *Drosophila*. *Cell* **153**, 797–811 (2013).
- Lemaître, B. & Miguel-Aliaga, I. The digestive tract of *Drosophila melanogaster*. *Annu. Rev. Genet.* **47**, 377–404 (2013).
- Kimura, K. I. & Truman, J. W. Postmetamorphic cell death in the nervous and muscular systems of *Drosophila melanogaster*. *J. Neurosci.* **10**, 403–411 (1990).
- Tissot, M. & Stocker, R. F. Metamorphosis in *Drosophila* and other insects: the fate of neurons throughout the stages. *Prog. Neurobiol.* **62**, 89–111 (2000).
- Blacher, E., Levy, M., Tatirovsky, E. & Elinav, E. Microbiome-modulated metabolites at the interface of host immunity. *J. Immunol.* **198**, 572–580 (2017).
- Breton, J. et al. Gut commensal *E. coli* proteins activate host satiety pathways following nutrient-induced bacterial growth. *Cell Metab.* **23**, 324–334 (2016).
- Mann, K., Gordon, M. D. & Scott, K. A pair of interneurons influences the choice between feeding and locomotion in *Drosophila*. *Neuron* **79**, 754–765 (2013).
- Wong, A. C.-N., Dobson, A. J. & Douglas, A. E. Gut microbiota dictates the metabolic response of *Drosophila* to diet. *J. Exp. Biol.* **217**, 1894–1901 (2014).
- Kim, E.-K., Park, Y. M., Lee, O. Y. & Lee, W.-J. Draft genome sequence of *Lactobacillus brevis* strain EW, a *Drosophila* gut pathobiont. *Genome Announc.* **1**, e00938-13 (2013).
- Martino, M. E. et al. Resequencing of the *Lactobacillus plantarum* strain WJL genome. *Genome Announc.* **3**, e01382-15 (2015).
- Baba, T. et al. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.008 (2006).
- Yamanaka, K. Purification, crystallization and properties of the D-xylose isomerase from *Lactobacillus brevis*. *Biochim. Biophys. Acta* **151**, 670–680 (1968).
- Ridley, E. V., Wong, A. C. N., Westmiller, S. & Douglas, A. E. Impact of the resident microbiota on the nutritional phenotype of *Drosophila melanogaster*. *PLoS ONE* **7**, e36765 (2012).
- Yang, Z. et al. Octopamine mediates starvation-induced hyperactivity in adult *Drosophila*. *Proc. Natl Acad. Sci. USA* **112**, 5219–5224 (2015).
- Chen, A. et al. Dispensable, redundant, complementary, and cooperative roles of dopamine, octopamine, and serotonin in *Drosophila melanogaster*. *Genetics* **193**, 159–176 (2013).
- Riemsperger, T. et al. Behavioral consequences of dopamine deficiency in the *Drosophila* central nervous system. *Proc. Natl Acad. Sci. USA* **108**, 834–839 (2011).
- Mithieux, G. et al. Portal sensing of intestinal gluconeogenesis is a mechanistic link in the diminution of food intake induced by diet protein. *Cell Metab.* **2**, 321–329 (2005).
- Hamada, F. N. et al. An internal thermal sensor controlling temperature preference in *Drosophila*. *Nature* **454**, 217–220 (2008).
- Roeder, T. Tyramine and octopamine: ruling behavior and metabolism. *Annu. Rev. Entomol.* **50**, 447–477 (2005).
- Crocker, A. & Sehgal, A. Octopamine regulates sleep in *Drosophila* through protein kinase A-dependent mechanisms. *J. Neurosci.* **28**, 9377–9385 (2008).
- Crocker, A., Shahidullah, M., Levitan, I. B. & Sehgal, A. Identification of a neural circuit that underlies the effects of octopamine on sleep:wake behavior. *Neuron* **65**, 670–681 (2010).
- Selcho, M., Pauls, D., El Jundi, B., Stocker, R. F. & Thum, A. S. The role of octopamine and tyramine in *Drosophila* larval locomotion. *J. Comp. Neurol.* **520**, 3764–3785 (2012).
- Saraswati, S., Fox, L. E., Soll, D. R. & Wu, C. F. Tyramine and octopamine have opposite effects on the locomotion of *Drosophila* larvae. *J. Neurobiol.* **58**, 425–441 (2004).
- Klaassen, L. W. & Kammer, A. E. Octopamine enhances neuromuscular transmission in developing and adult moths, *Manduca sexta*. *J. Neurobiol.* **16**, 227–243 (1985).
- Weisel-Eichler, A. & Libersat, F. Neuromodulation of flight initiation by octopamine in the cockroach *Periplaneta americana*. *J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol.* **179**, 103–112 (1996).
- Brembs, B., Christiansen, F., Pflüger, H. J. & Duch, C. Flight initiation and maintenance deficits in flies with genetically altered biogenic amine levels. *J. Neurosci.* **27**, 11122–11131 (2007).

44. van Breugel, F., Suver, M. P. & Dickinson, M. H. Octopaminergic modulation of the visual flight speed regulator of *Drosophila*. *J. Exp. Biol.* **217**, 1737–1744 (2014).
45. Han, D., Stein, D. & Stevens, L. M. Investigating the function of follicular subpopulations during *Drosophila* oogenesis through hormone-dependent enhancer-targeted cell ablation. *Development* **127**, 573–583 (2000).
46. Selkirk, J. et al. The *Drosophila* microbiome has a limited influence on sleep, activity, and courtship behaviors. *Sci. Rep.* **8**, 10646 (2018).
47. Nishino, R. et al. Commensal microbiota modulate murine behaviors in a strictly contamination-free environment confirmed by culture-based methods. *Neurogastroenterol. Motil.* **25**, 521–528 (2013).
48. Lendrum, J. E., Seebach, B., Klein, B. & Liu, S. Sleep and the gut microbiome: antibiotic-induced depletion of the gut microbiota reduces nocturnal sleep in mice. Preprint at <https://www.biorxiv.org/content/early/2017/10/05/199075> (2017).
49. Berridge, C. W. Noradrenergic modulation of arousal. *Brain Res. Rev.* **58**, 1–17 (2008).

Acknowledgements We thank H. Chu, G. Sharon, W.-L. Wu, J. K. Scarpa, E. D. Hooper and the Mazmanian laboratory for critiques; A. A. Aravin and K. Fejes Tóth for use of their laboratory space; D. J. Anderson, H. A. Lester, V. Gradinaru and M.-F. Chesselet for discussions; A. R. Sandoval, M. Meyerowitz and M. Smalley for technical support; Y. Garcia-Flores for administrative support; D. C. Hall for creating custom Python scripts; W.-J. Lee for the *L. brevis*^{EW}, *L. plantarum*^{WJ} and *Acetobacter pomorum* bacterial strains; the Yale Coli Genetic Stock Center for wild-type and mutant *E. coli* strains; M. H. Dickinson, D. J. Anderson, A. A. Aravin, and K. Fejes Tóth for fly lines; the GlycoAnalytics Core for help with carbohydrate analysis; and M. Fischbach and M. Funabashi for the pGID023 vector and advice. Imaging was performed

in the Biological Imaging Facility, with the support of the Caltech Beckman Institute and the Arnold and Mabel Beckman Foundation. C.E.S. was partially supported by the Center for Environmental Microbial Interactions at Caltech. This project was funded by grants from the NIH (NS085910) and the Heritage Medical Research Institute to S.K.M.

Reviewer information *Nature* thanks P. Bercik, C.-F. Wu and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.E.S. designed, performed and analysed most of the experiments. J.V. assisted with experimental design for biochemical analysis. I.B., Z.M., and S.M. assisted with gait analysis experiments. S.A. performed carbohydrate quantification. C.E.S. and S.K.M supervised the project. C.E.S. and S.K.M. wrote the manuscript with assistance from all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0634-9>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0634-9>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to C.E.S. or S.K.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Fly stocks and rearing. We obtained Canton-S (#64349), *Imd*^{-/-} (#55711), *Ti*^{-/-} (#30652), *UAS-dTrpA1* (#26264), *Tdc2-GAL4* (#52243), *T β h-GAL4* (#48332), *Th-GAL4* (#8488), *Ddc-GAL4* (#7009), *Gad1-GAL4* (#51630), *ChAT-GAL4* (#60317), *Elav-GAL4* (#46655), *UAS-T β h^{RNAi}* (#27667), *UAS-DTI* (#25039) and *pBPG4U-GAL4* (#68384) lines from Bloomington *Drosophila* Stock Center at Indiana University. Other fly stocks used were Oregon^R (kindly provided by A. A. Aravin and K. Fejes Tóth), *T β h^{M18}* (kindly provided by M. H. Dickinson)⁵⁰, *Tdc2^{R054}* and *tsh-GAL80* (both kindly provided by D. J. Anderson)^{51,52}. To minimize the effect of genetic background on behaviours, mutant fly lines were outcrossed for at least three generations onto a wild-type background.

Flies were cultured at 25 °C and 60% humidity on a 12-h light:12-h dark cycle and kept in vials containing fresh fly medium made at California Institute of Technology consisting of cornmeal, yeast, molasses, agar and p-hydroxy-benzoic acid methyl ester. Other dietary compositions used were created by altering this standard diet or the Nutri-Fly 'German Food' formula (Genesee Scientific) and were calculated using previously published nutritional data⁵³. Axenic flies were generated using standard methods^{18,54-58}. In brief, embryos from conventional flies were washed in bleach, ethanol and sterile PBS before being cultivated on fresh irradiated medium⁵⁴. Axenic stocks were maintained through the application of an irradiated diet supplemented with antibiotics (500 µg/ml ampicillin, Putney; 50 µg/ml tetracycline, Sigma; 200 µg/ml rifamycin, Sigma) for at least one generation. For experiments, virgin female flies were collected shortly after eclosion and placed at random into vials (10–15 flies per vial) containing irradiated medium without antibiotics. Vials were changed every 3–4 days using sterile methods. The antibiotic-supplemented diet was given to conventional flies shortly after eclosion to generate antibiotic-treated (ABX) flies. Both antibiotic-treated and axenic flies were tested for contaminants by plating animal lysates on Man, Rogosa and Sharpe (MRS, BD Biosciences), mannitol (25 g/l mannitol, Sigma; 5 g/l yeast extract, BD Biosciences; 3 g/l peptone, BD Biosciences), and Luria–Bertani (LB, BD Biosciences) nutrient agar plates.

Bacterial strains. *L. brevis*^{EW}, *L. plantarum*^{WJL} and *A. pomorum* were obtained from laboratory-reared flies in the laboratory of W.-J. Lee (Seoul National University)^{18,56,58}. *L. brevis*^{Bb14} (ATCC, #14869) and *L. brevis*^{P-2} (ATCC, #27305) were isolated from human faeces and fermented beverages, respectively. *E. coli*^{K12} (CGSC, #7636) was grown in LB broth and *E. coli*^{Atyra} (CGSC, #9131), *E. coli*^{AtypC} (CGSC, #10049), *E. coli*^{ΔmanX} (CGSC, #9511), *E. coli*^{ΔtreA} (CGSC, #9090), and *E. coli*^{ΔxylA} (CGSC, #10610)²⁸ were grown in LB broth supplemented with kanamycin (50 µg/ml). *L. brevis* and *L. plantarum* cultures were grown overnight in a standing 37 °C incubator in MRS broth (BD Biosciences). For mono-associations, fresh stationary phase bacterial cultures (OD₆₀₀ = 1.0, 40 µl) were added directly to fly vials. Associations with two bacteria were performed in a 1:1 mixture. For heat-killed bacteria experiments, fresh cultures of *L. brevis* (OD₆₀₀ = 1.0) were washed three times in sterile PBS, incubated at 100 °C for 30 min, and cooled to room temperature before being administered to flies. All treatments were supplied daily through application to the fly medium (40 µl) for 6 days following eclosion.

Bacterial supernatant preparations.

Cell-free supernatants (CFS) of specified bacterial strains were collected from bacterial cultures (OD₆₀₀ = 1.0) by centrifuging at 13,000g for 10 min and subsequent filtration through a 0.22-µm sterile filter (Millipore). CFS was dialysed in MilliQ water with a 3.5-kDa membrane (Thermo Scientific) overnight at 4 °C to generate *L. b* CFS and *L. p* CFS samples. Each of these treatments was supplied daily by application to the fly medium (40 µl) for 6 days following eclosion.

Heat and enzymatic treatment of *L. b* CFS. For heat-inactivation experiments, freshly prepared *L. b* CFS samples were incubated at 100 °C for 30 min and cooled to room temperature before being administered to flies. For proteinase K (PK) and trypsin (Typ) treatment, overnight dialysis of CFS was performed in Tris-HCl (pH 8 for PK and pH 8.5 for Typ) after which samples were treated with PK (100 µg/ml, Invitrogen) or Typ (0.05 µg/ml, Sigma) at 37 °C for 24 or 7 h, respectively. A proteinase inhibitor cocktail (Sigma) was added to stop the reaction and subsequently removed through overnight dialysis (Thermo Scientific) at 4 °C in MilliQ water. Aliquots of the samples were run on a 4–20% Tris-glycine gel (Invitrogen) to confirm protein cleavage. Controls followed the same protocol except for the addition of proteinase K or trypsin. For amylase digests, 20 µl of 100 mU/ml amylase (Sigma) was added to either freshly prepared *L. b* CFS or a PBS control for 30 min, and inhibited by lowering the pH to 4.5. Each of these treatments was supplied daily through application to the fly medium (40 µl) for 6 days following eclosion.

Production of His-tagged proteins (Xi* and Ai*).

An expression plasmid for the production of His-tagged Xi from *L. b* (Xi*) was constructed by amplification of its gene and cloning of the resulting PCR product into the pQE30 cloning vector (Qiagen) using SLIC ligation. The following primer sequences were used for the construct: 5'-CGCATCACCACCATCACGGATCTTA CTTGCTAACGCTATCGATGATGTAA-3' and 5'-GGGGTACCGAGCTCG CATCGGGATCATGACTGAAGAATACTGGAAAGGC-3'. The conformation of the resulting plasmid was verified and it was transformed into *E. coli* (Turbo,

NEB). This strain was then grown in LB medium containing ampicillin (100 µg/ml) and chloramphenicol (25 µg/ml) with shaking at 220 rpm at 37 °C for 1 h before the addition of 0.1 mM IPTG. After 4 h of shaking at 220 rpm at 37 °C, cells were pelleted and lysed using lysozyme (Sigma) and bead beating with matrix B beads (MP Biomedicals) for 45 s. Supernatant was collected after centrifugation and the Xi* protein purified through metal affinity purification under native conditions using HisPur Ni-NTA Spin Columns (Thermo Scientific). Protein purification was verified through western blot using an anti-6 × His tag antibody (Abcam) and quantified using a Pierce BCA Protein Assay kit (Thermo Scientific) after which protein was stored at -20 °C. Expression and purification of His-tagged L-arabinose isomerase from *L. b* (Ai*) was performed under the same conditions and the following primer sequences were used for the construct: 5'-GGGGTACCGAGCTCG CATGCGGATCATGTTATCAGTCCAGATTATGAATTTGG-3' and 5'-CGC ATCACCATCACCATCACGGATCCTTACTTGATGAACGCCCTTGTAT-3'. For EDTA treatment²⁹, purified Xi* was combined with 5 mM EDTA for 44 h at 4 °C and subsequently dialysed before administering to flies through application to the fly medium (40 µl) for 6 days following eclosion.

Generation of *xylA* deletion mutant (Δ*xylA*). Approximately 1-kb DNA segments flanking the region to be deleted were PCR-amplified using the following primers: 5'-ATTCCAATACTACCACTAGCAACGACATCCGTAAGT-3'; 5'-AATTGGAG CTCGGTACCCGGGGATCCACATCAGAATTGATCGCGGCAAC-3'; 5'-TCG TTGCTAGTGGTAGTATTGGAATCTAACAGATTCTTATCTTGTATG-3'; 5'-GCCTGCAGGTGCACTCTAGAGGATCCCGCAAGTCTAGTGCAGGCT-3'. The forward primers were designed using to be partially complementary at their 5' ends by 25 bp. The fused PCR product was cloned into the BamHI site of the *Lactobacilli* vector pGID023 and mobilized into *L. b*. Colonies selected for erythromycin (Erm) resistance, indicating integration of the vector into the host chromosome, were re-plated onto MRS + Erm and subsequently passaged over 5 days and plated onto MRS + Erm. Colonies selected for Erm resistance were passaged again in MRS alone over 3 days and plated on MRS. The resulting colonies were plated in replica on MRS and MRS + Erm. Erm-sensitive colonies were screened by PCR to distinguish wild-type revertants from strains with the desired mutation.

Drug treatments. Axenic flies were either left untreated or treated with *L. b* CFS or Xi* for 3 days after eclosion. After switching to new irradiated fly medium, groups of axenic flies were treated by application to the fly medium (40 µl) with octopamine (OA, 10 mg/ml, Sigma), tyramine (TA, 10 mg/ml, Sigma), L-dopa (1 mg/ml, Sigma) or mianserin (2 mg/ml) every day for 3 days before testing, similar to previously published methods^{33,37}.

Bacterial load quantification. Intestines dissected from surface-sterilized 7-day-old adult female flies were homogenized in sterile PBS with ~100 µl matrix D beads using a bead beater. Lysate dilutions in PBS were plated on MRS agar plates and enumerated after 24 h at 37 °C.

Locomotion assays. Locomotor behaviour was assayed by three previously established methods: the *Drosophila* activity monitoring system (DAMS, Trikinetics)^{59,60}, video-assisted tracking⁶¹⁻⁶³ and gait analysis⁶⁴.

Activity measurements. Seven-day-old individual female flies were cooled on ice for 1 min and transferred into individual vials (25 × 95 mm) containing standard irradiated medium. Tubes were then inserted and secured into *Drosophila* activity monitors (DAMS, Trikinetics) and kept in a fly incubator held at 25 °C. Flies were allowed to acclimate to the new environment for 1 day before testing and midline crossing was sampled every minute. Average daily activity was calculated from the 2 days tested and actograms were generated using Actogram⁶⁰. Sleep was defined as a 5 min bout of inactivity, as previously described⁶⁵.

Video-assisted tracking. Individual female flies were cooled on ice for 1 min before being introduced under sterile conditions into autoclaved arenas (3.5-cm diameter wells), which allowed free movement but restricted flight. After a 1-h acclimation period, arenas were placed onto a light box and recorded from above for a period of 10 min at 30 frames per s. All testing took place between ZT 0 and ZT 3 (ZT, Zeitgeber time; lights are turned on at ZT 0 and turned off at ZT 12) and both acclimation and testing occurred at 25 °C unless otherwise stated. Videos were processed using Ethovision software or the Caltech FlyTracker (<http://www.vision.caltech.edu/Tools/FlyTracker/>).

Bout analysis was performed using custom Python scripts (available upon request). The velocity curve was smoothed from the acquired video at 30 frames per s using a 15-s moving average window. A minimum walking speed of 0.25 mm/s was given, below which flies were moving but not walking ('pause bouts') and above which they were designated as walking ('walking bouts'). Lengths were measured as time between bout onset and offset.

Gait analysis. Experiments used an internally illuminated glass surface with frustrated total internal refraction to mark the flies' contact with the glass⁶⁴. The movement of the flies and their contact were recorded with a high-frame-rate camera, and videos were quantified using the FlyWalker software package. For further details of the parameters, see ref. ⁶⁴. All groups consisted of 7-day-old female flies and were tested at room temperature.

Feeding assays. Female flies were collected at the same time as described for locomotor assays. Flies were transferred regularly onto fresh food until day 7, upon which the flies were starved for 2 h and subsequently transferred for 30 min to irradiated standard fly medium dyed with FD&C Blue no. 1 (Sigma) at a final concentration of 0.5 g dye per 100 g food. Flies were allowed to feed on the food (3–4 biological replicates and 7 flies per replicate) at 25 °C after which they were decapitated and their bodies collected. Each replicate was homogenized in 150 µl of PBS/0.05% Triton X-100 and centrifuged at 5,000g for 1 min to remove debris. Absorbance for all groups was measured together at 630 nm and the amount of food consumed was estimated from a standard curve of the same dye solution. The manual feeding (MAFE) assay was performed as previously described^{66,67}. In brief, individual flies were introduced into a 200-µl pipette tip, which was cut to expose the proboscis. Flies were first water-satiated and presented with 100 mM sucrose delivered in a fine graduated capillary (VWR). After flies were unresponsive to 10 food stimuli, the assay was terminated and the total volume of food was calculated.

Measurement of life span. Adult female flies were transferred under sterile conditions to irradiated fly medium every 4–5 days. Survival in three or more independent cohorts containing 15–25 flies each was monitored over time.

Apoptosis assay. Midguts from 7-day-old female flies were dissected in PBS containing 0.1% Triton X-100 and the apoptosis assay was performed as previously described^{18,56}. The percentage of apoptotic cells was determined by dividing the number of apoptotic cells by the total number of cells in each section and multiplying by 100.

Measurement of carbohydrate levels. Fly (5 flies per sample) and fly medium (0.1 g per sample) samples were homogenized in TE buffer (10 mM Tris, pH = 8, 1 mM EDTA) using a bead beater for 45 s followed by centrifugation at 7,000g for 3 min. The supernatant was heat-treated for 30 min at 72 °C before being stored at –80 °C before subsequent clean-up steps before running on high-performance anion exchange chromatography with pulsed amperometric detection.

One hundred microlitres of fly or fly medium homogenate in TE buffer was diluted with 200 µl UltraPure distilled water (Invitrogen) and sonicated to obtain a uniform solution. Samples were centrifuged at 2,000 rpm for 15 s to precipitate insoluble material. One hundred microlitres of the sample was filtered through a pre-washed Pall Nanosep 3K Omega centrifugal device (MWCO 3KDa, Sigma-Aldrich) for 15 min at 14,000 rpm and 7 °C. The filtrate was dried on Speed Vac. The dry sample was reconstituted in 300 µl UltraPure water and loaded onto a pre-washed Dionex OnGuard IIH 1cc cartridge. The flow through and 2 × 1 ml elution with Ultrapure water were collected in the same tube and lyophilized.

Monosaccharide analysis was done using a Dionex CarboPac PA1 column (4 × 250 mm) with PA1 guard column (4 × 50 mm); flow rate, 1 ml/min; pulsed amperometric detection with gold electrode. The elution gradient was as follows: 0–20 min, 19 mM sodium hydroxide; 20–50 min, 0–212.5 mM sodium acetate gradient with 19 mM sodium hydroxide; 50–65 min, 212.5 mM sodium acetate with 19 mM sodium hydroxide; 65–68 min, 212.5–0 mM sodium acetate with 19 mM sodium hydroxide; 68–85 min, 19 mM sodium hydroxide.

Trehalose, arabinose, galactose, glucose, mannose, xylose, fructose, ribose, sucrose and xylulose were used as standards. The monosaccharides were assigned based on the retention time and quantified using Chromeleon 6.8 chromatography data system software. In Extended Data Fig. 7f, g, measurements of trehalose levels were performed following the same isolation procedure and subsequently processed using a Trehalose Assay Kit (Megazyme) according to the manufacturer's instructions.

For experiments in which flies were treated with trehalose, arabinose or ribose, groups of axenic or axenic flies previously treated with *Xi** were given trehalose, arabinose or ribose (10 mg/ml, Sigma) through application to the fly medium (40 µl) every day for 3 days before testing.

RNA isolation and quantitative real-time PCR. Heads (20 flies per sample) or decapitated bodies (5 flies per sample) were dissected on ice and immediately processed using an Arcturus PicoPure RNA isolation kit (Applied Biosystems) or a standard TRIzol-chloroform protocol (Thermo Fisher). One microgram of RNA was reverse transcribed using iScript cDNA Synthesis Kit, according to the manufacturer's protocol, (Bio-Rad) and diluted to 10 ng/µl based on the input concentration of total RNA.

Previously published primer pairs were used to target immune-related gene transcripts^{18,68}. Other primer sequences used include *Tdc2* (F: GGTCTGCCGG ACCACTTTC, R: CACTCCGATGCCGAAGTCTG), *Tβh* (F: GCTTATCCGA CACAAAGCTGC, R: GAAAGCATTCTGCAAGTGGAA), *Ddc* (F: TGGGAT GAGCACACCATCTTG, R: GTAGAAGGAAATCAAACCCCTCG), *Tph* (also known as *Trh*) (F: TGTTTCGCCAAGGATTCTG, R: CACCAAGGTT TATGTCATGCTTCT). All primers were synthesized by Integrated DNA Technologies. Real-time PCR for the house-keeping genes *Rp49* and *RpL32* was

used to ensure that input RNA was equal among all samples. Real-time PCR was performed on cDNA using an ABI PRISM 7900 HT system (Thermo Fisher) according to the manufacturer's instructions.

Data reporting and statistical analysis. No statistical methods were used to pre-determine sample size. Sample size was based on previous literature in the field and experimenters were not blinded as almost all data acquisition and analysis were automated. After eclosion, virgin female flies with the same genotype were sorted into groups of 10–15 flies per vial at random. All flies in each vial were given the same treatment regime. For each experiment, the experimental and control flies were collected, treated and tested at the same time. A Mann–Whitney *U* test or Kruskal–Wallis test and Dunn's post hoc test were used for statistical analysis of behavioural data and carbohydrate analysis. Comparisons with more than one variant were first analysed using two-way ANOVA. An unpaired two-sided Student's *t*-test or a one-way ANOVA followed by a Bonferroni post hoc test were used for statistical analysis of quantitative RT–PCR results and CFU analysis. All statistical analysis was performed using Prism Software (GraphPad, version 7). *P* values are indicated as follows: ****P* < 0.0001; ***P* < 0.001; ***P* < 0.01; and **P* < 0.05. See Supplementary Information for more details on statistical tests and exact *P* values for each figure. For boxplots, lower and upper whiskers represent 1.5 × IQR of the lower and upper quartiles, respectively; boxes indicate lower quartile, median, and upper quartile, from bottom to top. When all points are shown, whiskers represent range and boxes indicate lower quartile, median, and upper quartile, from bottom to top. Bar graphs are presented as mean ± s.e.m.

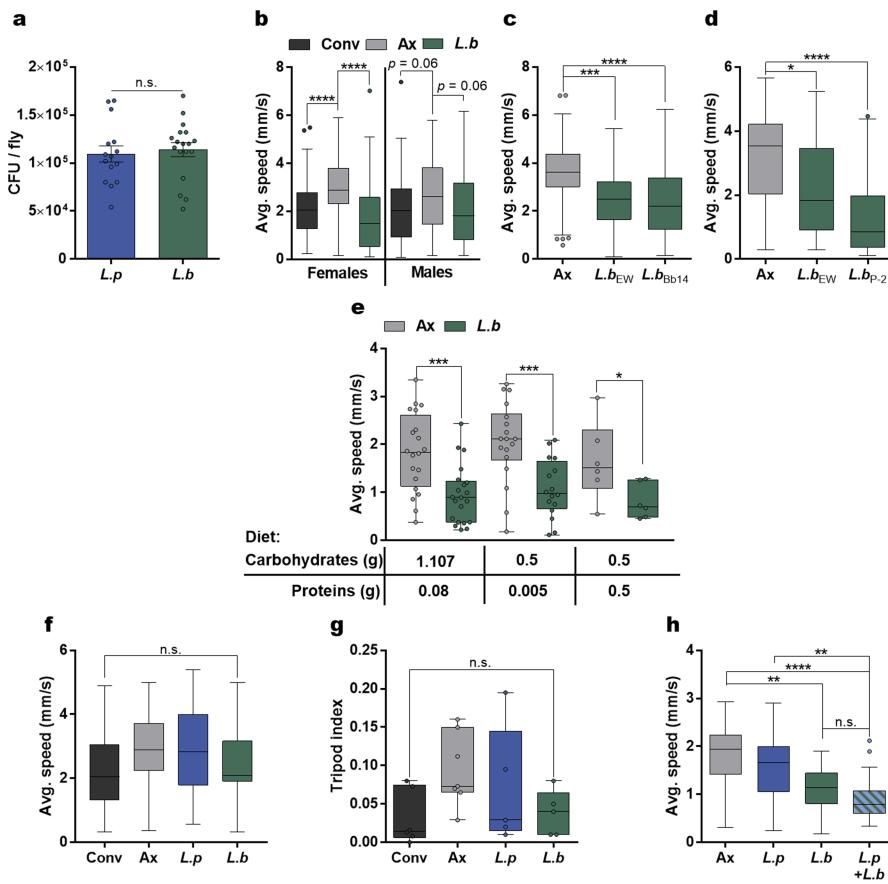
Code availability. Custom code for bout analysis is available from the corresponding authors upon request.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

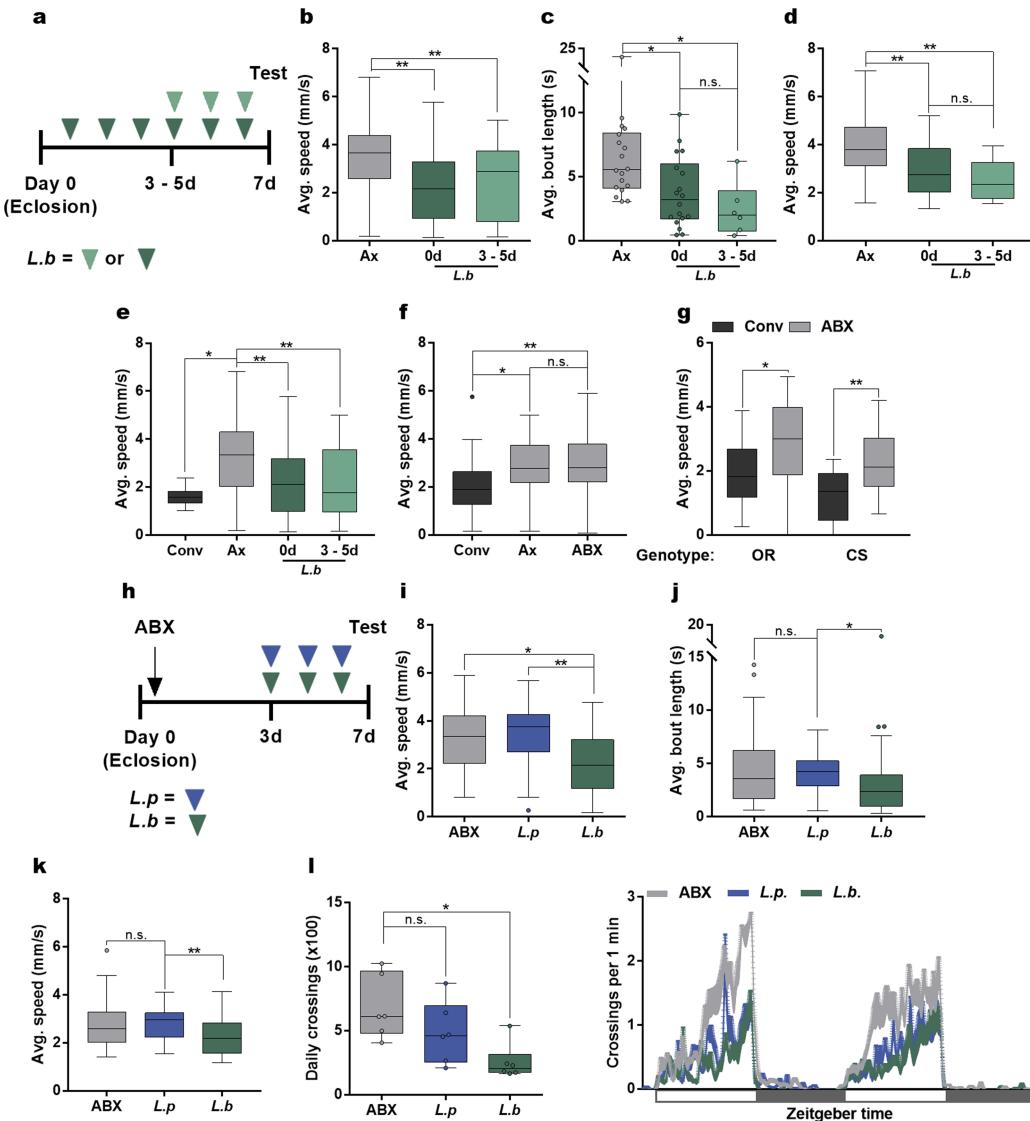
All datasets generated are available from the corresponding authors upon request.

- Monastirioti, M., Linn, C. E. Jr & White, K. Characterization of *Drosophila* tyramine beta-hydroxylase gene and isolation of mutant flies lacking octopamine. *J. Neurosci.* **16**, 3900–3911 (1996).
- Clyne, J. D. & Miesenböck, G. Sex-specific control and tuning of the pattern generator for courtship song in *Drosophila*. *Cell* **133**, 354–363 (2008).
- Shiga, Y., Tanaka-Matakatsu, M. & Hayashi, S. A nuclear GFP/β-galactosidase fusion protein as a marker for morphogenesis in living *Drosophila*. *Dev. Growth Differ.* **38**, 99–106 (1996).
- Lee, W. C. & Micchelli, C. A. Development and characterization of a chemically defined food for *Drosophila*. *PLoS ONE* **8**, e67308 (2013).
- Brummel, T., Ching, A., Seroude, L., Simon, A. F. & Benzer, S. *Drosophila* lifespan enhancement by exogenous bacteria. *Proc. Natl Acad. Sci. USA* **101**, 12974–12979 (2004).
- Ren, C., Webster, P., Finkel, S. E. & Tower, J. Increased internal and external bacterial load during *Drosophila* aging without life-span trade-off. *Cell Metab.* **6**, 144–152 (2007).
- Ryu, J.-H. et al. Innate immune homeostasis by the homeobox gene caudal and commensal-gut mutualism in *Drosophila*. *Science* **319**, 777–782 (2008).
- Storelli, G. et al. *Lactobacillus plantarum* promotes *Drosophila* systemic growth by modulating hormonal signals through TOR-dependent nutrient sensing. *Cell Metab.* **14**, 403–414 (2011).
- Shin, C. et al. *Drosophila* microbiome modulates host developmental and metabolic homeostasis via insulin signaling. *Science* **334**, 670–674 (2011).
- Chiu, J. C., Low, K. H., Pike, D. H., Yildirim, E. & Edery, I. Assaying locomotor activity to study circadian rhythms and sleep parameters in *Drosophila*. *J. Vis. Exp.* **43**, 2157 (2010).
- Schmid, B., Helfrich-Förster, C. & Yoshii, T. A new ImageJ plug-in "ActogramJ" for chronobiological analyses. *J. Biol. Rhythms* **26**, 464–467 (2011).
- Wolf, F. W., Rodan, A. R., Tsai, L. T.-Y. & Heberlein, U. High-resolution analysis of ethanol-induced locomotor stimulation in *Drosophila*. *J. Neurosci.* **22**, 11035–11044 (2002).
- Simon, J. C. & Dickinson, M. H. A new chamber for studying the behavior of *Drosophila*. *PLoS ONE* **5**, e8793 (2010).
- White, K. E., Humphrey, D. M. & Hirth, F. The dopaminergic system in the aging brain of *Drosophila*. *Front. Neurosci.* **4**, 205 (2010).
- Mendes, C. S., Bartos, I., Akay, T., Márka, S. & Marin, R. S. Quantification of gait parameters in freely walking wild type and sensory deprived *Drosophila melanogaster*. *eLife* **2**, e00231 (2013).
- Shaw, P. J., Cirelli, C., Greenspan, R. J. & Tononi, G. Correlates of sleep and waking in *Drosophila melanogaster*. *Science* **287**, 1834–1837 (2000).
- Yu, Y. et al. Regulation of starvation-induced hyperactivity by insulin and glucagon signaling in adult *Drosophila*. *eLife* **5**, e15693 (2016).
- Qi, W. et al. A quantitative feeding assay in adult *Drosophila* reveals rapid modulation of food ingestion by its nutritional value. *Mol. Brain* **8**, 87 (2015).
- Chakrabarti, S., Poidevin, M. & Lemaitre, B. The *Drosophila* MAPK p38c regulates oxidative stress and lipid homeostasis in the intestine. *PLoS Genet.* **10**, e1004659 (2014).



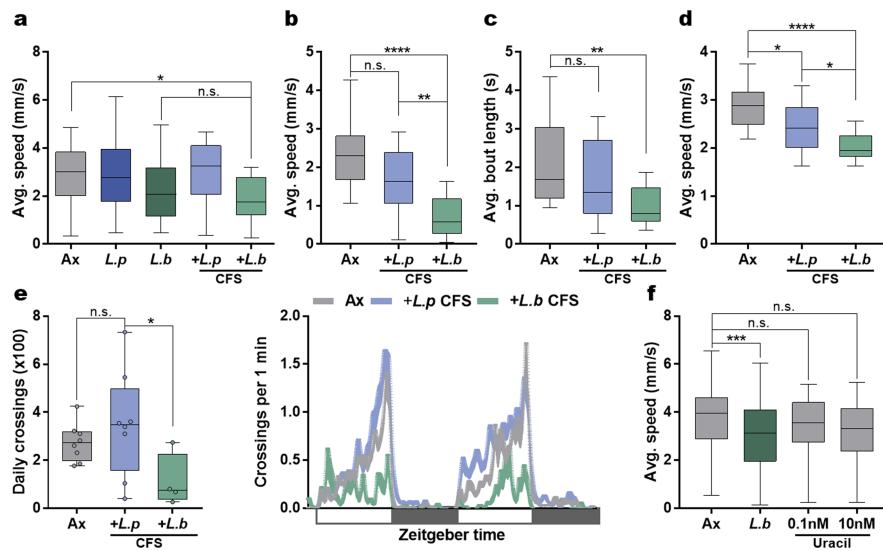
Extended Data Fig. 1 | Effects of colonization level, bacterial strain, and host diet on *L. brevis* modulation of locomotion. **a**, Colony-forming units (CFU) per individual fly (mean \pm s.e.m.) for *L.p* or *L.b* mono-associated flies. *L.p*, $n=15$; *L.b*, $n=18$. **b**, Average speed of Conv, Ax and *L.b* mono-associated female or male flies. Females: Conv, $n=90$; Ax, $n=92$; *L.b*, $n=89$; Males: Conv, $n=100$; Ax, $n=100$; *L.b*, $n=95$. **c, d**, Average speed of Ax flies or flies mono-associated with *L.b* strains EW, Bb14 or P-2. **c**, Ax, $n=58$; *L.b* EW, $n=57$; *L.b* Bb14, $n=57$. **d**, Ax, $n=45$; *L.b* EW, $n=28$; *L.b* P-2, $n=42$. **e**, Average speed of Ax or *L.b* mono-associated flies raised on different diet compositions from eclosion until day 7. Diet 1 (left): Ax, $n=20$; *L.b*, $n=21$; diet 2 (middle): Ax, $n=18$; *L.b*, $n=16$; diet 3 (right): Ax, $n=6$; *L.b*, $n=6$. **f**, Average speed during

walking bouts for Conv, Ax, *L.p* and *L.b* groups. Conv, $n=23$; Ax, $n=35$; *L.p*, $n=22$; *L.b*, $n=22$. **g**, Tripod index for Conv, Ax, *L.p* and *L.b* groups. Conv, $n=6$; Ax, $n=7$; *L.p*, $n=5$; *L.b*, $n=5$. **h**, Average speed of Ax flies or flies mono-associated with *L.p* or *L.b* alone or in combination (1:1). Ax, $n=18$; *L.p*, $n=24$; *L.b*, $n=24$; *L.p* + *L.b*, $n=24$. Box-and-whisker plots show median and IQR; whiskers show either $1.5 \times$ IQR of the lower and upper quartiles or range. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$. Specific P values are in the Supplementary Information. Unpaired Student's t -test (**a**), Kruskal–Wallis and Dunn's (**b–d**, **f–h**), or Mann–Whitney U (**e**) post hoc tests were used for statistical analysis. Data are representative of at least three independent trials for each experiment.

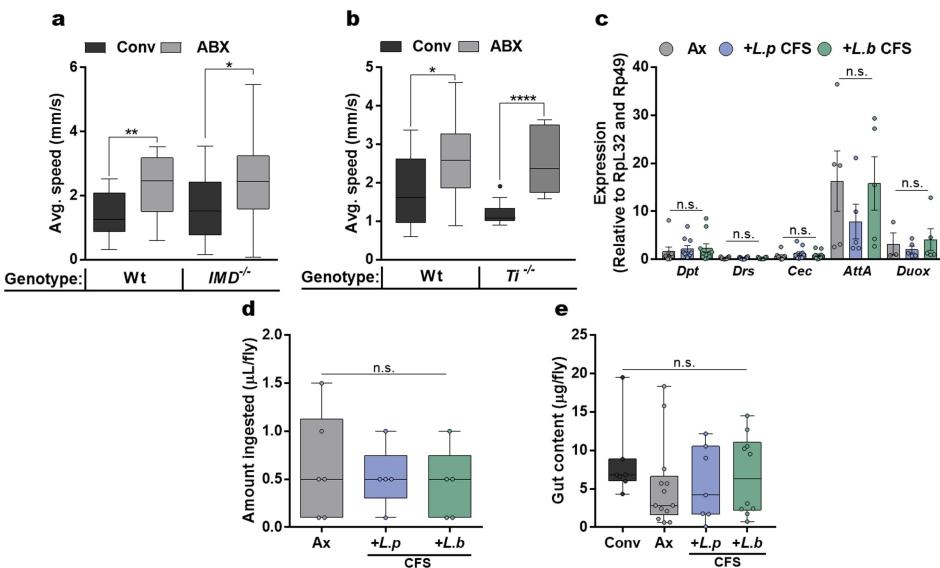


Extended Data Fig. 2 | Post-eclosion microbial signals decrease host locomotion. **a**, Experimental design (b–e) in which Ax flies were associated with *L.b.* either directly after (day 0, dark green arrows) or 3–5 days after (light green arrows) eclosion. **b–d**, Average speed (b), average bout length (c) and average speed during walking bouts (d) of Ax flies and flies mono-associated with *L.b.* at either day 0 or day 3–5. **b**, Ax, $n=46$; *L.b.* 0 d, $n=47$; *L.b.* 3–5 d, $n=43$. **c**, Ax, $n=18$; *L.b.* 0 d, $n=18$; *L.b.* 3–5 d, $n=6$. **d**, Ax, $n=36$; *L.b.* 0 d, $n=36$; *L.b.* 3–5 d, $n=12$. **e**, Average speed of Conv flies, Ax flies and flies mono-associated with *L.b.* at either day 0 or day 3–5. Conv, $n=11$; Ax, $n=53$; *L.b.* 0 d, $n=53$; *L.b.* 3–5 d, $n=52$. **f**, Average speed of Conv, Ax and Conv flies treated with antibiotics for 3 days after eclosion (ABX). Conv, $n=32$; Ax, $n=36$; ABX, $n=36$. **g**, Average speed of Oregon^R (OR) and Canton-S (CS) Conv flies and Conv flies treated with antibiotics for 3 days after eclosion (ABX). OR: Conv, $n=20$; ABX, $n=22$; CS: Conv, $n=12$; ABX, $n=17$. **h**, Experimental design (i–l) in which conventionally reared flies were treated with antibiotics (ABX, black arrow) for 3 days following eclosion. All flies were subsequently placed on irradiated medium either without supplementation (ABX) or associated with *L.p.* (blue arrows) or *L.b.* (green arrows) for the 3 days before testing. **i–k**, Average speed (i), average bout length (j) and average speed during walking bouts (k) calculated for ABX, *L.p.*- and *L.b*-associated flies. **i**, ABX, $n=29$; *L.p.*, $n=24$; *L.b.*, $n=35$. **j**, ABX, $n=36$; *L.p.*, $n=30$; *L.b.*, $n=35$. **k**, ABX, $n=42$; *L.p.*, $n=30$; *L.b.*, $n=35$. **l**, Daily activity of ABX, *L.p.* and, *L.b.* groups (virgin female Oregon^R flies) over a 2-day 12 h light:12 h dark cycle period, starting at time 0. White boxes represent lights on and grey boxes represent lights off. $n=6$ per condition. Box-and-whisker plots show median and IQR; whiskers show either $1.5 \times$ IQR of the lower and upper quartiles or range. * $P < 0.05$, ** $P < 0.01$. Specific P values are in the Supplementary Information. Kruskal-Wallis and Dunn's (b–f, i–l) or Mann-Whitney U (g) post hoc tests were used for statistical analysis. Data are representative of at least two independent trials for each experiment.

arrow) for 3 days following eclosion. All flies were subsequently placed on irradiated medium either without supplementation (ABX) or associated with *L.p.* (blue arrows) or *L.b.* (green arrows) for the 3 days before testing. **i–k**, Average speed (i), average bout length (j) and average speed during walking bouts (k) calculated for ABX, *L.p.*- and *L.b*-associated flies. **i**, ABX, $n=29$; *L.p.*, $n=24$; *L.b.*, $n=35$. **j**, ABX, $n=36$; *L.p.*, $n=30$; *L.b.*, $n=35$. **k**, ABX, $n=42$; *L.p.*, $n=30$; *L.b.*, $n=35$. **l**, Daily activity of ABX, *L.p.* and, *L.b.* groups (virgin female Oregon^R flies) over a 2-day 12 h light:12 h dark cycle period, starting at time 0. White boxes represent lights on and grey boxes represent lights off. $n=6$ per condition. Box-and-whisker plots show median and IQR; whiskers show either $1.5 \times$ IQR of the lower and upper quartiles or range. * $P < 0.05$, ** $P < 0.01$. Specific P values are in the Supplementary Information. Kruskal-Wallis and Dunn's (b–f, i–l) or Mann-Whitney U (g) post hoc tests were used for statistical analysis. Data are representative of at least two independent trials for each experiment.

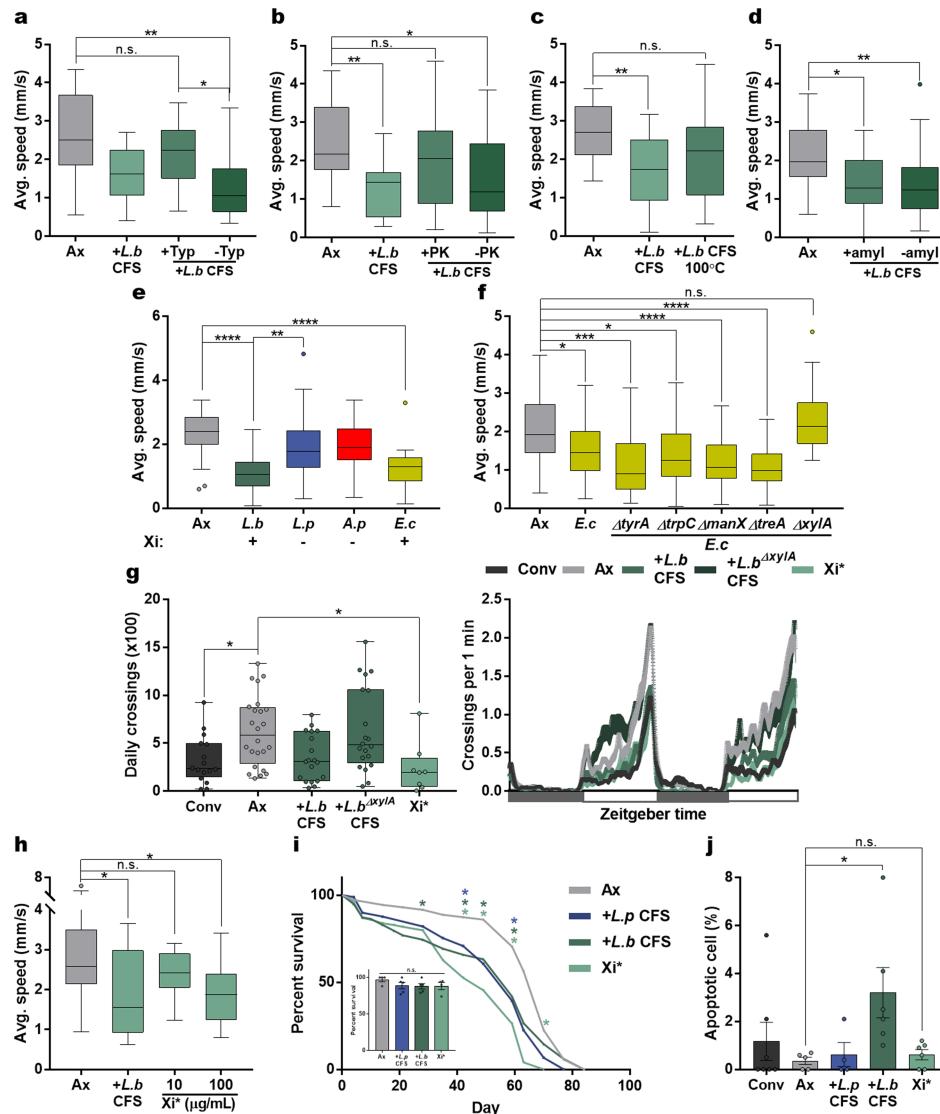


Extended Data Fig. 3 | Bacterial-derived products from *L. brevis* alter locomotion. **a**, Average speed of Ax flies, *L.p* or *L.b* mono-associated flies, and Ax flies treated with CFS from *L.p* or *L.b*. Ax, $n = 45$; *L.p*, $n = 17$; *L.b*, $n = 42$; *L.p* CFS, $n = 17$; *L.b* CFS, $n = 16$. **b–e**, Average speed (**b**), average bout length (**c**), average speed during walking bouts (**d**) and daily activity (**e**) of Ax flies and Ax virgin female Oregon^R flies treated with CFS from *L.p* or *L.b*. White boxes represent lights on and grey boxes represent lights off. **b**, Ax, $n = 23$; *L.p* CFS, $n = 20$; *L.b* CFS, $n = 20$. **c**, Ax, $n = 23$; *L.p* CFS, $n = 20$; *L.b* CFS, $n = 17$. **d**, Ax, $n = 22$; *L.p* CFS, $n = 21$; *L.b* CFS, $n = 17$. **e**, Ax, $n = 8$; *L.p* CFS, $n = 8$; *L.b* CFS, $n = 4$. **f**, Average speed of Ax, *L.b* mono-associated and Ax uracil-treated flies. Ax, $n = 96$; *L.b*, $n = 88$; 0.1 nM uracil, $n = 41$; 10 nM uracil, $n = 18$. Box-and-whisker plots show median and IQR; whiskers show either $1.5 \times$ IQR of the lower and upper quartiles or range. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$. Specific P values are in the Supplementary Information. Kruskal–Wallis and Dunn's post hoc tests were used for statistical analysis. Data are representative of at least two independent trials for each experiment.



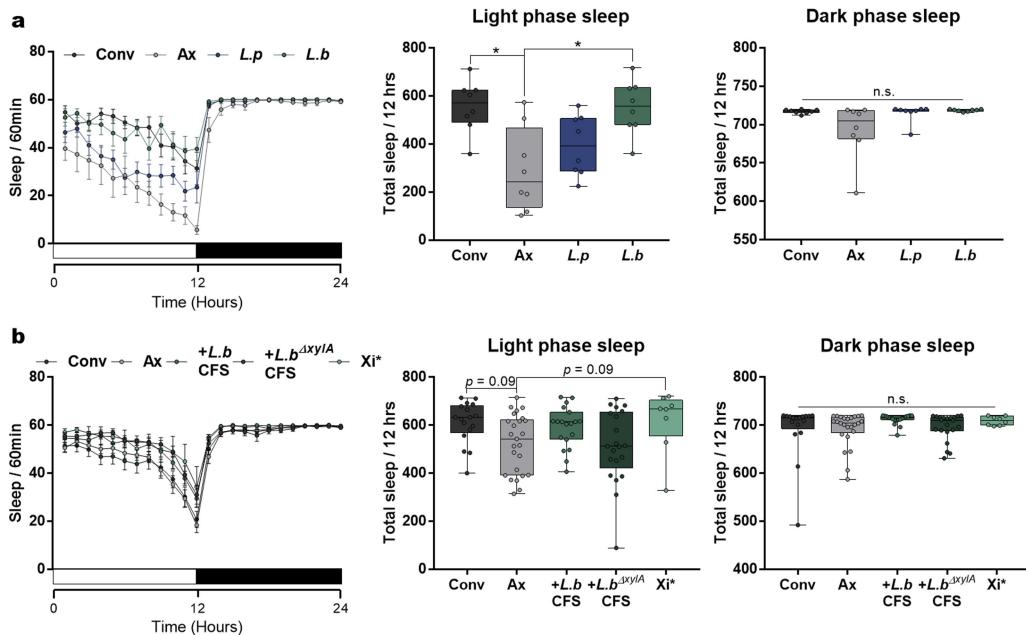
Extended Data Fig. 4 | Locomotor phenotypes are independent of food intake, anti-microbial peptides, and the immune deficiency (IMD) and Toll pathways. **a**, Average speed of wild-type background (Oregon^R, Wt) and *Imd*^{-/-} flies placed on either medium alone or medium supplemented with antibiotics (ABX) following eclosion. Wt: Conv, $n = 16$; ABX, $n = 17$; *IMD*^{-/-}: Conv, $n = 24$; ABX, $n = 25$. **b**, Average speed of wild-type background (Canton-S, Wt) and *Ti*^{-/-} flies placed on either medium alone or medium supplemented with antibiotics (ABX) following eclosion. Wt: Conv, $n = 15$; ABX, $n = 17$; *Ti*^{-/-}: Conv, $n = 10$; ABX, $n = 11$. **c**, qRT-PCR of immune-related transcripts (mean \pm s.e.m.) in Ax and Ax *L.p* or *L.b* CFS-treated flies. *Dpt* (also known as *DptA*): Ax, $n = 8$; *L.p* CFS, $n = 10$; *L.b* CFS, $n = 10$; *Drs*: Ax, $n = 10$; *L.p* CFS, $n = 10$; *L.b* CFS, $n = 10$; *Cec* (also known as *CecA1*): Ax, $n = 8$; *L.p* CFS, $n = 10$; *L.b* CFS, $n = 10$; *AttA*: Ax,

$n = 5$; *L.p* CFS, $n = 5$; *L.b* CFS, $n = 5$; *Duo*: Ax, $n = 3$; *L.p* CFS, $n = 5$; *L.b* CFS, $n = 5$. **d**, Amount ingested by Ax and Ax *L.p* or *L.b* CFS-treated flies over 10 trials during MAFFE assay. Ax, $n = 6$; *L.p* CFS, $n = 5$; *L.b* CFS, $n = 6$. **e**, Intestinal content measured through supplementing the diet of Conv, Ax, and *L.p*- or *L.b*-CFS-treated Ax flies with blue food dye. Conv, $n = 7$; Ax, $n = 13$; *L.p* CFS, $n = 7$; *L.b* CFS, $n = 10$. Box-and-whisker plots show median and IQR; whiskers show either $1.5 \times$ IQR of the lower and upper quartiles or range. $*P < 0.05$, $**P < 0.01$, $****P < 0.0001$. Specific P values are in the Supplementary Information. Mann-Whitney U (**a**, **b**), one-way ANOVA and Bonferroni (**c**), and Kruskal-Wallis and Dunn's (**d**, **e**) post hoc tests were used for statistical analysis. Data are representative of at least two independent trials for each experiment. *Dpt*, diptericin; *Drs*, drosomycin; *Cec*, cecropin; *AttA*, attacin-A; *Duo*, dual oxidase.



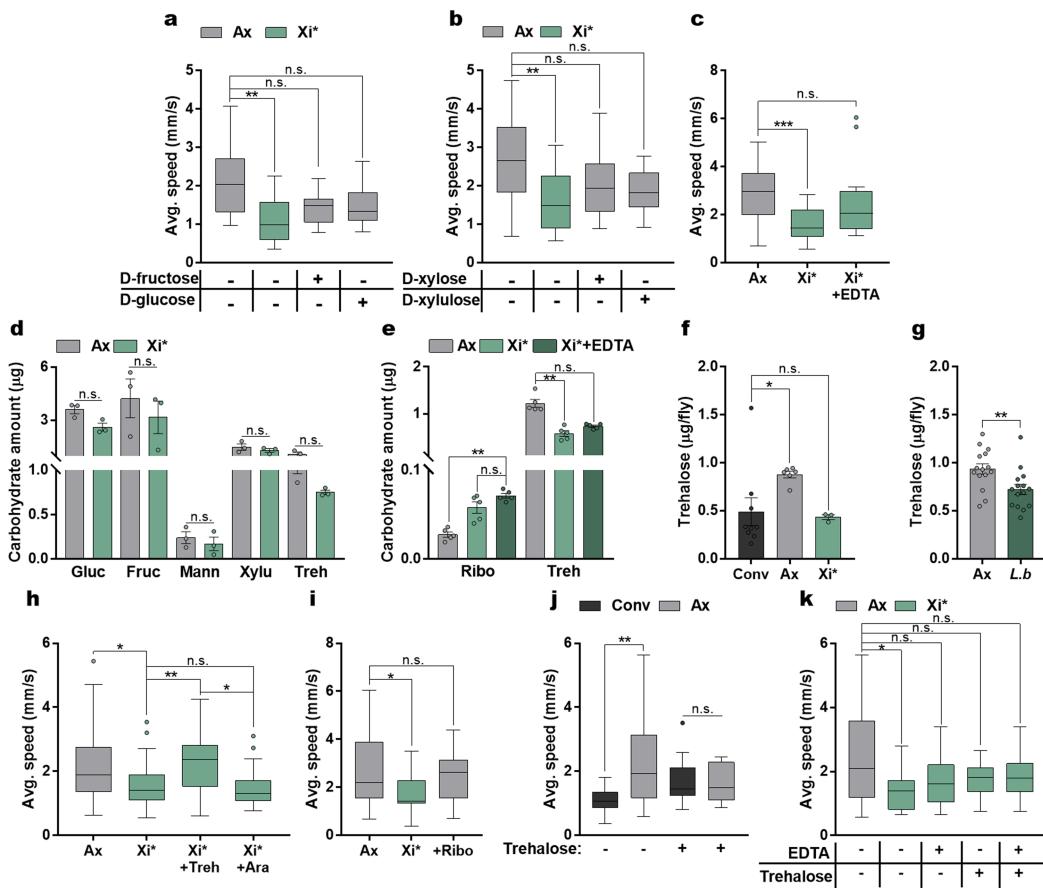
Extended Data Fig. 5 | Modulation of locomotion by the bacterial enzyme, xylose isomerase. **a–c**, Average speed of Ax flies or Ax flies treated with unaltered, protease-treated (Typ, trypsin; PK, proteinase-K) or heat-treated (100 °C) *L. b* CFS. **a**, Ax, $n = 18$; *L. b* CFS, $n = 18$; +Typ, $n = 17$; -Typ, $n = 17$. **b**, Ax, $n = 23$; *L. b* CFS, $n = 18$; +PK, $n = 23$; -PK, $n = 23$. **c**, $n = 18$. **d**, Average speed of Ax flies treated with amylase-treated PBS (Ax), amylase-treated *L. b* CFS (+ amyl *L. b* CFS) or unaltered *L. b* CFS (-amyl *L. b* CFS). Ax, $n = 30$; +amyl, $n = 17$; -amyl, $n = 30$. **e**, Average speed of Ax flies or flies mono-associated with *L. b*, *L. p*, *A. pomorum* (*A. p*), or *E. coli* (*E. c*). Below shows the presence (+) or absence (-) of *E. coli* based on NCBI Blastn (*xylA* locus) and Blastp (*Xi*) results. Ax, $n = 30$; *L. b*, $n = 30$; *L. p*, $n = 29$; *A. p*, $n = 30$; *E. c*, $n = 18$. **f**, Average speed of Ax flies and flies mono-associated with either WT *E. coli* or single gene knockout strains of *E. coli* (Δ *tyrA*, Δ *trpC*, Δ *manX*, Δ *treA*, Δ *xylA*). Ax, $n = 65$; *E. c*, $n = 52$; *E. c* Δ *tyrA*, $n = 18$; *E. c* Δ *trpC*, $n = 17$; *E. c* Δ *manX*, $n = 45$; *E. c* Δ *treA*, $n = 46$; *E. c* Δ *xylA*, $n = 20$. **g**, Daily activity of Conv, Ax and Ax virgin female Oregon^R flies treated with *L. b* CFS, *L. b* Δ *xylA* CFS or *Xi** over a two-day 12 h light:12 h dark cycle. **h**, Average speed (mm/s) of Ax flies and Ax flies treated with *L. b* CFS or *Xi**. Ax, $n = 16$; *L. b* CFS, $n = 11$; 10 µg/ml *Xi**, $n = 12$; 100 µg/ml *Xi**, $n = 14$. **i**, Lifespan measurements for Ax flies and Ax flies treated with *L. p* CFS, *L. b* CFS, or *Xi**. Asterisks represent significance at the time point measured by Kruskal–Wallis and Dunn's post hoc test. Inset image shows survival at day 7 (mean \pm s.e.m.). Ax, $n = 4$ groups; *L. p* CFS, $n = 5$ groups; *L. b* CFS, $n = 5$ groups; *Xi**, $n = 4$ groups. **j**, Percentage of apoptotic cells (mean \pm s.e.m.) in the intestines of Conv flies, Ax flies and Ax flies treated with *L. p* CFS, *L. b* CFS or *Xi**. Conv, $n = 7$; Ax, $n = 5$; *L. p* CFS, $n = 4$; *L. b* CFS, $n = 6$; *Xi**, $n = 6$. Box-and-whisker plots show median and IQR; whiskers show either 1.5 \times IQR of the lower and upper quartiles or range. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$. Specific P values are in the Supplementary Information. Kruskal–Wallis and Dunn's (a–i) or log-rank (i) post hoc tests were used for statistical analysis. Data are representative of at least two independent trials for each experiment.

period, starting at time 0. White boxes represent lights on and grey boxes represent lights off. Conv, $n = 16$; Ax, $n = 24$; *L. b* CFS, $n = 19$; *L. b* Δ *xylA* CFS, $n = 20$; *Xi**, $n = 8$. **h**, Average speed of Ax flies and Ax flies treated with *L. b* CFS or *Xi**. Ax, $n = 16$; *L. b* CFS, $n = 11$; 10 µg/ml *Xi**, $n = 12$; 100 µg/ml *Xi**, $n = 14$. **i**, Lifespan measurements for Ax flies and Ax flies treated with *L. p* CFS, *L. b* CFS, or *Xi**. Asterisks represent significance at the time point measured by Kruskal–Wallis and Dunn's post hoc test. Inset image shows survival at day 7 (mean \pm s.e.m.). Ax, $n = 4$ groups; *L. p* CFS, $n = 5$ groups; *L. b* CFS, $n = 5$ groups; *Xi**, $n = 4$ groups. **j**, Percentage of apoptotic cells (mean \pm s.e.m.) in the intestines of Conv flies, Ax flies and Ax flies treated with *L. p* CFS, *L. b* CFS or *Xi**. Conv, $n = 7$; Ax, $n = 5$; *L. p* CFS, $n = 4$; *L. b* CFS, $n = 6$; *Xi**, $n = 6$. Box-and-whisker plots show median and IQR; whiskers show either 1.5 \times IQR of the lower and upper quartiles or range. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$. Specific P values are in the Supplementary Information. Kruskal–Wallis and Dunn's (a–i) or log-rank (i) post hoc tests were used for statistical analysis. Data are representative of at least two independent trials for each experiment.



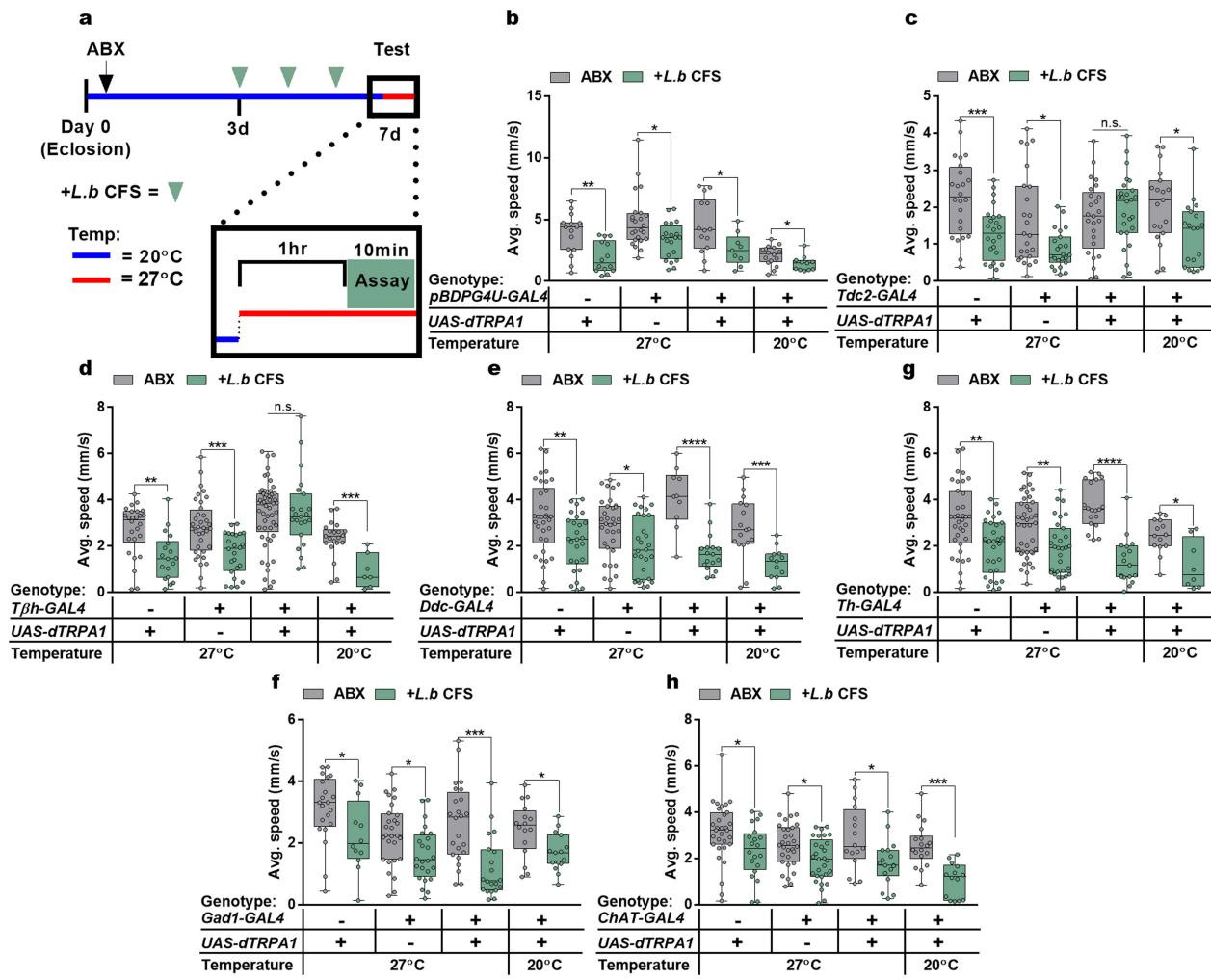
Extended Data Fig. 6 | Sleep analysis for mono-colonized flies and flies treated with bacterial factors. a. Twenty-four-hour sleep profiles (mean \pm s.e.m.) of Conv, Ax, *L.p*- and *L.b*-colonized virgin female Oregon^R flies with the number of sleep bouts in 60-min time windows and total sleep in the light or dark phase. $n=8$ flies per condition. **b.** Twenty-four-hour sleep profiles (mean \pm s.e.m.) of Conv, Ax, *L.b* CFS, *L.b*^{ΔxyIA} CFS and *Xi** treated Ax virgin female Oregon^R flies with the number of sleep bouts

in 60-min time windows and total sleep in the light or dark phase. Conv, $n=17$; Ax, $n=25$; *L.b* CFS, $n=19$; *L.b*^{ΔxyIA} CFS, $n=21$; *Xi**, $n=8$. Box-and-whisker plots show median and IQR; whiskers show range. $^*P<0.05$. Specific *P* values are in the Supplementary Information. Kruskal–Wallis and Dunn's post hoc tests were used for statistical analysis. Data are representative of at least two independent trials for each experiment.



Extended Data Fig. 7 | Xylose isomerase activity and key carbohydrates are involved in Xi-mediated changes in locomotion. **a, b,** Average speed of Ax flies and Ax flies treated with Xi^* or 100 $\mu\text{g}/\text{ml}$ of D-fructose, D-glucose, D-xylose or D-xylulose. **a,** Ax, $n = 16$; Xi^* , $n = 13$; D-fructose, $n = 13$; D-glucose, $n = 15$. **b,** Ax, $n = 26$; Xi^* , $n = 21$; D-xylose, $n = 22$; D-xylulose, $n = 18$. **c,** Average speed of Ax flies and Ax flies treated with Xi^* or Xi^* inactivated by 5 mM EDTA. Ax, $n = 21$; Xi^* , $n = 16$; $\text{Xi}^* + \text{EDTA}$, $n = 18$. **d,** Carbohydrate levels (mean \pm s.e.m.) in Ax and Xi^* -treated fly medium. Each sample is from 0.1 g fly medium and represents a separate vial. $n = 3$ samples per condition. **e,** Carbohydrate levels (mean \pm s.e.m.) in Ax, Xi^* , and EDTA-treated Xi^* flies. Each sample contains five flies. $n = 5$ samples per condition. **f,** Trehalose levels (mean \pm s.e.m.) in Conv, Ax, and Xi^* -treated flies. Conv, $n = 9$ samples; Ax, $n = 6$ samples; Xi^* , $n = 3$ samples. **g,** Trehalose levels (mean \pm s.e.m.) in Ax and *L.b*-colonized flies. $n = 15$ samples per condition. **h,** Average speed of Ax and Xi^* -treated flies supplemented with either trehalose (Treh, 10 mg/ml) or arabinose (Ara, 10 mg/ml) for 3 days before testing.

Ax, $n = 40$; Xi^* , $n = 40$; $\text{Xi}^* + \text{Treh}$, $n = 39$; $\text{Xi}^* + \text{Ara}$, $n = 18$. **i,** Average speed of Ax flies and Xi^* - or ribose (Ribo, 10 mg/ml)-treated flies. Ax, $n = 29$; Xi^* , $n = 25$; Ribo, $n = 12$. **j,** Average speed of Conv and Ax flies supplemented with trehalose (Treh, 10 mg/ml) for 3 days before testing. Conv, $n = 15$; Ax, $n = 22$; Conv + Treh, $n = 18$; Ax + Treh, $n = 15$. **k,** Average speed of Ax and Xi^* or EDTA-treated Xi^* Ax flies subsequently left untreated or supplemented with trehalose (Treh, 10 mg/ml) for 3 days before testing. Ax, $n = 27$; Xi , $n = 19$; $\text{Xi} + \text{EDTA}$, $n = 24$; $\text{Xi} + \text{Treh}$, $n = 19$; $\text{Xi} + \text{EDTA} + \text{Treh}$, $n = 25$. Box-and-whisker plots show median and IQR; whiskers show $1.5 \times$ IQR of the lower and upper quartiles. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. Specific P values are in the Supplementary Information. Kruskal-Wallis and Dunn's (a-c, e, f, h-k) or Mann-Whitney U (d, g) post hoc tests were used for statistical analysis. Data are representative of at least two independent trials for each experiment. Gluc, glucose; Fruc, fructose; Mann, mannose; Xylu, xylulose; Treh, trehalose; Ribo, ribose.

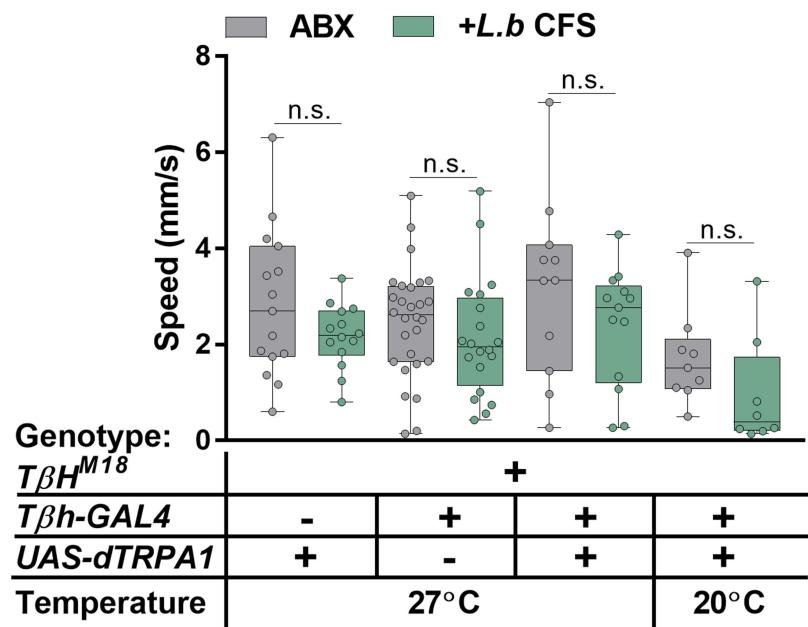


Extended Data Fig. 8 | Thermogenetic activation of neuromodulator GAL4 lines. **a**, Experimental design in which Conv flies (Canton-S) were treated with antibiotics (ABX, black arrow) for 3 days following eclosion. All flies were subsequently placed on irradiated medium either without supplementation or treated with *L.b* CFS (green arrows) for 3 days. One hour before and during testing, flies were either exposed to 27°C (red line) to facilitate thermogenetic activation or kept at 20°C (blue line).

b-h, Average speed of flies previously treated with antibiotics and subsequently left untreated (ABX) or treated with *L.b* CFS for 3 days before testing. **b**, UAS: ABX, $n = 15$; *L.b* CFS, $n = 14$; GAL4: ABX, $n = 24$; *L.b* CFS, $n = 20$; GAL4 > UAS (27°C): ABX, $n = 14$; *L.b* CFS, $n = 9$; GAL4 > UAS (20°C): ABX, $n = 16$; *L.b* CFS, $n = 11$. **c**, UAS: ABX, $n = 24$; *L.b* CFS, $n = 24$; GAL4: ABX, $n = 24$; *L.b* CFS, $n = 23$; GAL4 > UAS (27°C): ABX, $n = 25$; *L.b* CFS, $n = 26$; GAL4 > UAS (20°C): ABX, $n = 19$; *L.b* CFS, $n = 19$. **d**, UAS: ABX, $n = 26$; *L.b* CFS, $n = 18$; GAL4: ABX, $n = 36$; *L.b* CFS, $n = 24$; GAL4 > UAS (27°C): ABX, $n = 53$; *L.b* CFS, $n = 23$; GAL4 > UAS

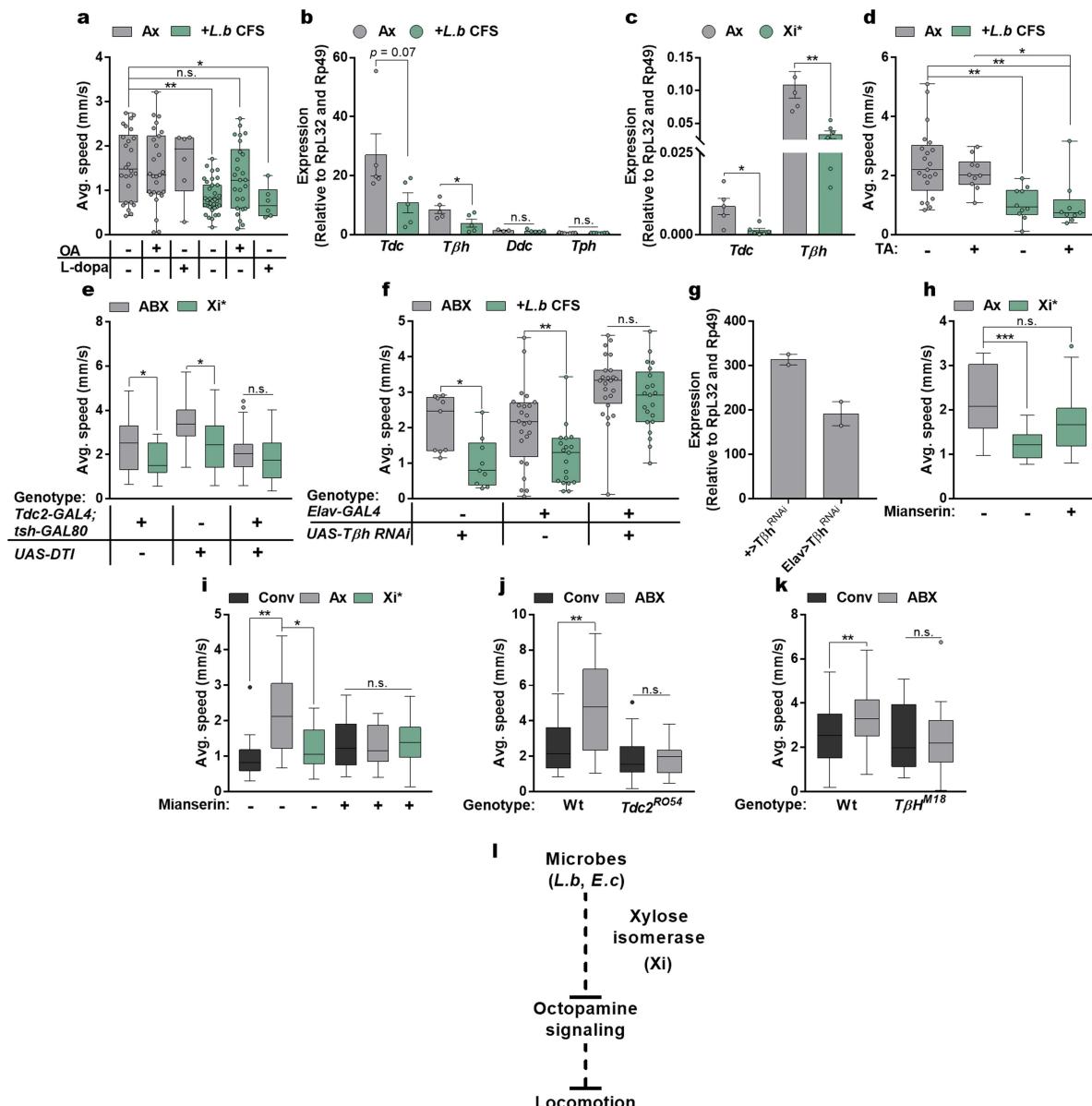
(20°C): ABX, $n = 21$; *L.b* CFS, $n = 7$. **e**, UAS: ABX, $n = 34$; *L.b* CFS, $n = 26$; GAL4: ABX, $n = 34$; *L.b* CFS, $n = 28$; GAL4 > UAS (27°C): ABX, $n = 10$; *L.b* CFS, $n = 17$; GAL4 > UAS (20°C): ABX, $n = 17$; *L.b* CFS, $n = 13$.

f, UAS: ABX, $n = 36$; *L.b* CFS, $n = 30$; GAL4: ABX, $n = 40$; *L.b* CFS, $n = 31$; GAL4 > UAS (27°C): ABX, $n = 19$; *L.b* CFS, $n = 17$; GAL4 > UAS (20°C): ABX, $n = 14$; *L.b* CFS, $n = 8$. **g**, UAS: ABX, $n = 21$; *L.b* CFS, $n = 12$; GAL4: ABX, $n = 28$; *L.b* CFS, $n = 24$; GAL4 > UAS (27°C): ABX, $n = 24$; *L.b* CFS, $n = 20$; GAL4 > UAS (20°C): ABX, $n = 16$; *L.b* CFS, $n = 15$. **h**, UAS: ABX, $n = 31$; *L.b* CFS, $n = 20$; GAL4: ABX, $n = 31$; *L.b* CFS, $n = 29$; GAL4 > UAS (27°C): ABX, $n = 16$; *L.b* CFS, $n = 17$; GAL4 > UAS (20°C): ABX, $n = 18$; *L.b* CFS, $n = 14$. Box-and-whisker plots show median and IQR; whiskers show range. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$. Specific P values are in the Supplementary Information. Mann-Whitney U post hoc tests following a two-way ANOVA were used for statistical analysis. Data are representative of at least two independent trials for each experiment.



Extended Data Fig. 9 | Activation of octopaminergic neurons in flies carrying a null allele for T β h ($T\beta h^{M18}$). Average speed of flies previously treated with antibiotics and subsequently left untreated (ABX) or treated with *L.b* CFS for 3 days before testing. UAS: ABX, $n = 15$; *L.b* CFS, $n = 14$; GAL4: ABX, $n = 28$; *L.b* CFS, $n = 20$; GAL4> UAS (27°C): ABX, $n = 11$;

L.b CFS, $n = 13$; GAL4> UAS (20°C): ABX, $n = 9$; *L.b* CFS, $n = 8$. Box-and-whisker plots show median and IQR; whiskers show range. Specific P values are in the Supplementary Information. Mann-Whitney U post hoc tests following a two-way ANOVA were used for statistical analysis. Data are representative of at least two independent trials.



Extended Data Fig. 10 | Octopamine mediates *L. brevis*- and *Xi*-induced changes in locomotion. **a**, Average speed of Ax and *L. b* CFS-treated Ax flies left untreated or supplemented with octopamine (OA, 10 mg/ml) or L-dopa (1 mg/ml) for 3 days. Ax, $n = 26$; Ax + OA, $n = 27$; Ax + L-dopa, $n = 6$; *L. b* CFS, $n = 35$; *L. b* CFS + OA, $n = 26$; *L. b* CFS + L-dopa, $n = 6$. **b**, RT-qPCR (mean \pm s.e.m.) for transcripts from heads of Ax and *L. b* CFS-treated Ax flies. *Tdc2*: $n = 5$; *Tβh*, $n = 5$; *Ddc*: Ax, $n = 3$; *L. b* CFS, $n = 5$; *Tph*: $n = 7$. **c**, qRT-PCR (mean \pm s.e.m.) for transcripts from heads of Ax or *Xi**-treated Ax flies. Ax, $n = 5$ samples; *Xi**, $n = 6$ samples. **d**, Average speed of Ax and *L. b* CFS-treated Ax flies left untreated or supplemented with tyramine (TA, 10 mg/ml) for 3 days. Ax, $n = 21$; Ax + TA, $n = 10$; *L. b* CFS, $n = 10$; *L. b* CFS + TA, $n = 9$. **e**, Average speed of control lines and flies expressing *DTI* in octopaminergic and tyraminergic neurons outside the ventral nerve cord. All flies were previously treated with antibiotics and subsequently left untreated (ABX) or treated with *Xi** for 3 days before testing. GAL4;GAL80: Ax, $n = 25$; *Xi**, $n = 18$; UAS: Ax, $n = 26$; *Xi**, $n = 21$; GAL4> UAS: Ax, $n = 39$; *Xi**, $n = 23$. **f**, Average speed of control lines and flies expressing *Tβh* RNAi in all neurons. All flies were previously treated with antibiotics and subsequently left untreated (ABX) or treated with *L. b* CFS for 3 days before testing. UAS: $n = 9$; GAL4: Ax, $n = 24$; *L. b* CFS, $n = 19$; GAL4> UAS: Ax, $n = 24$; *L. b* CFS, $n = 21$. **g**, *Tβh* mRNA measured from heads of flies previously treated with antibiotics.

Error bars represent range. $n = 2$ samples per condition. **h**, Average speed of Ax and *Xi**-treated Ax flies left untreated or supplemented with mianserin (Mian; 2 mg/ml) for 3 days. Ax, $n = 14$; *Xi**, $n = 15$; *Xi** + Mian, $n = 15$. **i**, Average speed of Conv, Ax and *Xi**-treated Ax flies left untreated or supplemented with mianserin (2 mg/ml) for 3 days. Conv, $n = 13$; Ax, $n = 28$; *Xi**, $n = 24$; Conv + Mian, $n = 27$; Ax + Mian, $n = 22$; *Xi** + Mian, $n = 22$. **j**, Average speed of wild-type background (w+, Wt) and *Tdc2*-null mutants (*Tdc2^{RO54}*) either left untreated or after treatment with antibiotics for 3 days following eclosion. Wt Conv, $n = 13$; Wt ABX, $n = 21$; *Tdc2^{RO54}* Conv, $n = 28$; *Tdc2^{RO54}* ABX, $n = 34$. **k**, Average speed of wild-type background (Canton-S, Wt) and *Tβh*-null mutants (*Tβh^{M18}*) either left untreated or after treatment with antibiotics for 3 days following eclosion. Wt Conv, $n = 38$; Wt ABX, $n = 42$; *Tβh^{M18}* Conv, $n = 25$; *Tβh^{M18}* ABX, $n = 33$. **l**, Model of bacterial modulation of host locomotion. Box-and-whisker plots show median and IQR; whiskers show either 1.5 \times IQR of the lower and upper quartiles or range. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Specific P values are in the Supplementary Information. Kruskal-Wallis and Dunn's (**a**, **d**, **h**, **i**), unpaired Student's *t*-test (**b**, **c**) or Mann-Whitney *U* (**e**, **f**, **j**, **k**) post hoc tests were used for statistical analysis. Data are representative of at least two independent trials for each experiment. *Tdc*, tyrosine decarboxylase; *Tβh*, tyramine beta-hydroxylase; *Ddc*, DOPA decarboxylase; *Tph*, tryptophan hydroxylase.

Distinct proteostasis circuits cooperate in nuclear and cytoplasmic protein quality control

Rahul S. Samant^{1,*}, Christine M. Livingston^{1,3*}, Emily M. Sontag¹ & Judith Frydman^{1,2,*}

Protein misfolding is linked to a wide array of human disorders, including Alzheimer's disease, Parkinson's disease and type II diabetes^{1,2}. Protective cellular protein quality control (PQC) mechanisms have evolved to selectively recognize misfolded proteins and limit their toxic effects^{3–9}, thus contributing to the maintenance of the proteome (proteostasis). Here we examine how molecular chaperones and the ubiquitin–proteasome system cooperate to recognize and promote the clearance of soluble misfolded proteins. Using a panel of PQC substrates with distinct characteristics and localizations, we define distinct chaperone and ubiquitination circuitries that execute quality control in the cytoplasm and nucleus. In the cytoplasm, proteasomal degradation of misfolded proteins requires tagging with mixed lysine 48 (K48)- and lysine 11 (K11)-linked ubiquitin chains. A distinct combination of E3 ubiquitin ligases and specific chaperones is required to achieve each type of linkage-specific ubiquitination. In the nucleus, however, proteasomal degradation of misfolded proteins requires only K48-linked ubiquitin chains, and is thus independent of K11-specific ligases and chaperones. The distinct ubiquitin codes for nuclear and cytoplasmic PQC appear to be linked to the function of the ubiquilin protein Dsk2, which is specifically required to clear nuclear misfolded proteins. Our work defines the principles of cytoplasmic and nuclear PQC as distinct, involving combinatorial recognition by defined sets of cooperating chaperones and E3 ligases. A better understanding of how these organelle-specific PQC requirements implement proteome integrity has implications for our understanding of diseases linked to impaired protein clearance and proteostasis dysfunction.

Misfolded proteins, arising during biogenesis or through proteotoxic damage, are highly toxic; they accumulate in distinct regions (puncta) within cells^{6–9} and form aggregates that are associated with neurodegenerative diseases¹. Misfolded proteins must be cleared: the process involves the cooperation of chaperones and components of the ubiquitin–proteasome system (UPS)^{3–5} (Fig. 1a), but is poorly understood.

To better understand the PQC of soluble proteins, we used a panel of substrates that reflects different types of misfolding, including two temperature-sensitive proteins (Ubc9^{ts} and luciferase^{ts}), a protein (the Von Hippel–Lindau tumour suppressor, VHL) that cannot fold without its oligomeric partners elongin B and C, and translocation-defective carboxypeptidase yscY (CPY^{*}) that lacks its signal sequence (CPY[‡])^{6,7,9,10}. We found that terminally misfolded VHL or CPY[‡] conjugated to green fluorescent protein (GFP) was cleared within 1 hour, with only around 10% of cells containing GFP-positive puncta (Fig. 1b, c). Blocking proteasomal degradation with the proteasome inhibitor bortezomib led to the accumulation of these proteins in GFP-positive puncta. Temperature-sensitive proteins, such as Ubc9^{ts}–GFP, are diffuse when folded at 30 °C, but upon misfolding at 37 °C are also degraded through the proteasome or accumulate in puncta (Extended Data Fig. 1a)^{6,7}. Of note, native wild-type Ubc9 is soluble at either temperature.

To identify the E3 ubiquitin (Ub) ligases involved in PQC, we expressed GFP–VHL in a library of 41 *Saccharomyces cerevisiae* single-deletion strains comprising most non-essential E3 ubiquitin ligases (Extended Data Fig. 1b), including E3s that have previously been implicated in PQC (Fig. 1d). No single deletion caused a notable increase in puncta formation compared with wild-type cells, suggesting that no single E3 is essential for PQC of this substrate.

Given that E3 ligases can have redundant functions¹¹, we next deleted pairs of E3s previously implicated in PQC. The cytoplasmic E3 Ubr1 and the nuclear E3 San1 cooperate in the clearance of some cytoplasmic proteins^{11,12}. The E3s Hrd1, which is anchored in the endoplasmic reticulum (ER) membrane, and Doa10, also localized to the ER and to the inner nuclear membrane¹³, trigger ER-associated protein degradation (ERAD)¹⁴. We found that both $\Delta ubr1\Delta san1$ and $\Delta doa10\Delta hrd1$ strains abrogated the degradation of all PQC substrates to the same extent as proteasome inhibition (see, for example, the results of cycloheximide chase in Fig. 1e, f and Extended Data Fig. 1c). Notably, all PQC substrates accumulated in the same puncta in these strains (Fig. 1g and Extended Data Fig. 1d). Therefore, multiple misfolded proteins use the same E3 systems for proteasomal degradation, and are sequestered together in the same PQC compartments in the absence of these E3 systems.

To determine why the deletion of either E3 pair stabilized our PQC substrates similarly, we tested the effect of doubly deleting all possible combinations of these four E3 ligases—Ubr1, San1, Doa10 and Hrd1 (Fig. 1h and Extended Data Fig. 1e). Puncta formation and cycloheximide chase assays showed that only specific combinations of deletions abrogate clearance. Strong stabilization was observed in $\Delta ubr1\Delta san1$ and $\Delta doa10\Delta hrd1$ strains. A moderate effect was found with $\Delta doa10\Delta ubr1$ and $\Delta hrd1\Delta san1$. Strikingly, $\Delta doa10\Delta san1$ and $\Delta hrd1\Delta ubr1$ had no effect on PQC, suggesting that these pairs of ligases—Doa10/San1 and Hrd1/Ubr1—provide parallel, optimal combinations for PQC clearance. This E3 circuit logic appeared to be general, as it operated for all substrates tested at 30 °C and 37 °C. Thus PQC clearance requires parallel pathways of specific pairs of E3 ligases, combining one of the soluble E3s (either Ubr1 or San1; blue in Fig. 1i) and one of the membrane-bound E3s (either Doa10 or Hrd1; red). Overexpressing an E3 in any of the double-deletion strains rescued clearance only in those strains deleted for that particular E3 (Extended Data Fig. 1f). Therefore, E3 function is not interchangeable, even at higher expression levels, and PQC requires specific E3 combinations. The functional cooperation between E3 ligases may involve a physical complex, as immunoprecipitation experiments show that San1 associates with Doa10 but not with Hrd1 (Extended Data Fig. 2). We were unable to co-immunoprecipitate Ubr1 with Hrd1 (or Doa10), so these ligases may bind transiently, or cooperate functionally through separate complexes.

To examine whether the PQC defects observed above were linked to impaired ubiquitination, we expressed a terminally misfolded Flag-tagged VHL mutant (L158P; ref. ¹⁰) in either wild-type cells or the panel of double E3 deletions. VHL was immunoprecipitated under harsh denaturing conditions, followed by anti-ubiquitin immunoblot

¹Department of Biology, Stanford University, Stanford, CA, USA. ²Department of Genetics, Stanford University, Stanford, CA, USA. ³Present address: Janssen Research and Development, Spring House, PA, USA. *e-mail: rsamant@stanford.edu; christine.marie.livingston@gmail.com; jfrydman@stanford.edu

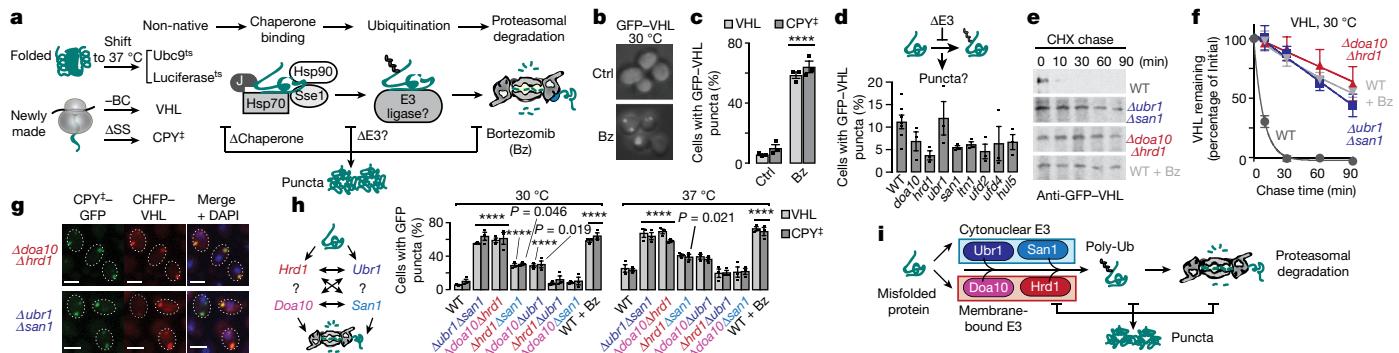


Fig. 1 | Combined deletion of certain E3 ligases impairs the clearance of misfolded proteins. **a**, Proteins can become misfolded for different reasons; in our experiments (involving non-native proteins), the temperature-sensitive proteins Ubc9^{ts} or luciferase^{ts} misfold at temperatures of 37 °C; VHL misfolds in the absence of elongin BC (BC); and CPY⁺ misfolds because it lacks its signal sequence (SS). Conserved pathways dispose of these aberrant proteins, with initial recognition by chaperones of the Hsp70 and Hsp90 families, J-domain proteins and Sse1; ubiquitination by one or more E3 ligases; and targeting for proteasomal degradation. Blocking any step (for instance, with the proteasome inhibitor bortezomib (Bz) or by omitting E3s or chaperones) triggers the sequestration of misfolded proteins into puncta. **b, c**, VHL and CPY⁺ form puncta upon proteasome inhibition. Wild-type cells expressing galactose-inducible GFP-VHL or CPY⁺-GFP were shifted to glucose media with 50 μM Bz or vehicle control (Ctrl) for 1 h to shut off expression. Fixed cells were imaged by fluorescence microscopy. **c**, Percentage of cells with GFP-VHL puncta. **d**, Deleting individual E3s implicated in PQC does not increase puncta formation. Experiment performed as in panel **b**, but using strains with endogenous deletions of E3s. WT, wild type. **e, f**, Misfolded proteins are stabilized in $\Delta ubr1\Delta san1$ and $\Delta doa10\Delta hrd1$

detection (Fig. 2a). In wild-type cells, polyubiquitin chains were clearly attached to VHL. In the $\Delta ubr1\Delta san1$ and $\Delta doa10\Delta hrd1$ cells, VHL ubiquitination was markedly reduced, indicating that these E3 pairs promote VHL clearance through ubiquitination. No clear differences in VHL ubiquitination were observed in any other strains. A similar strategy examined how E3 deletions specifically affect VHL tagging with a canonical proteasomal signal, K48-linked Ub (in which the lysine 48 residue of one ubiquitin molecule is joined to the carboxyl terminus of the next ubiquitin in the chain, and so on) (Fig. 2b)⁵. Surprisingly, K48-linked VHL ubiquitination was strongly reduced in $\Delta ubr1\Delta san1$ but not in $\Delta doa10\Delta hrd1$ cells. Conversely, the alternative proteasome-targeting signal, K11-linked Ub^{5,15–18}, was unaffected in $\Delta ubr1\Delta san1$ cells but was abrogated in $\Delta doa10\Delta hrd1$ cells. Note that misfolded VHL was ubiquitinated with both K48-linked and K11-linked ubiquitin chains in wild-type cells. A quantitative ubiquitin-linkage-specific ELISA assay for K11- or K48-linked chains confirmed that Ubr1 and San1 mediate VHL tagging with K48-Ub chains, whereas Doa10 and Hrd1 promote K11-linked ubiquitination (Extended Data Fig. 3a, b). Upon proteasome inhibition, the VHL puncta co-localized with both K11- and K48-linked ubiquitin (Fig. 2c), supporting the conclusion that VHL itself is modified with ubiquitin chains containing both K48 and K11 linkages.

We next examined PQC in cells expressing a single copy of ubiquitin that carries individual lysine to arginine (K-to-R) mutations, and is therefore unable to form specific linkages (Fig. 2d). Ub^{K48R} is lethal¹⁹, and could not be tested. VHL was degraded in all other K-to-R variants except in Ub^{K11R} cells, where VHL clearance was impaired, leading to its accumulation in puncta (Fig. 2e, f). Importantly, K11-Ub linkages were not generally required for proteasomal clearance, given that the non-misfolded cytoplasmic substrates Ub-R-GFP and Ub^{G76V}-GFP were degraded normally in Ub^{K11R} cells (Extended Data Fig. 4). Consistent with our data implicating ER-resident E3 ligases in K11-Ub tagging of soluble PQC substrates, the Ub^{K11R} strain has been shown to be sensitive to ER stress¹⁶.

strains: **e**, cycloheximide (CHX) chase and immunoblot were used to assess the stability of GFP-VHL in E3 double-deletion strains, or following 50 μM Bz treatment; **f**, densitometric quantification of bands relative to $t = 0$ (mean \pm s.e.m. from three biologically independent experiments). **g**, Multiple misfolded proteins are sequestered in the same subcellular location. Experiment performed as in panel **b**, but in strains co-expressing VHL with CPY⁺. Images represent more than 100 cells from each of three biologically independent experiments. Scale bars represent 2 μm. **h**, Deleting certain E3 pairs increases puncta formation. Experiment performed as in panel **b**, but using strains with endogenous deletions of E3 pairs. On the right, cells were shifted to 37 °C for the shut-off. **i**, Clearance of misfolded proteins requires a soluble E3, Ubr1 or San1, and a membrane-bound E3, Doa10 or Hrd1. **c, d, h**, 300 cells were counted per condition. Bars represent mean \pm s.e.m. from three biologically independent experiments, except for the wild type in panel **d** (seven biologically independent experiments). Statistically significant differences versus wild type, calculated by two-tailed Student's *t*-test (**c**), or one-way analysis of variance (ANOVA) plus Dunnett's multiple comparisons test (**d, h**), are indicated with the adjusted *P* value, or with *** for $P < 0.0001$.

We next examined whether misfolded VHL is tagged with mixed K48- and K11-linked ubiquitin chains. Sequential double immunoprecipitations first isolated Flag-tagged VHL and then isolated either K48- or K11-linked chains (Extended Data Fig. 5a). The presence of K11- or K48-linked chains in either flow-through or bound fractions was then detected by immunoblot. All of the K11- and K48-linked chains were present only in the bound fractions of each immunoprecipitate (Extended Data Fig. 5b), showing that misfolded VHL is tagged with both K11- and K48-linked ubiquitin chains. Additionally, an antibody that recognizes branched K11/K48 linkages, with K11- and K48-linked chains extending from the same ubiquitin molecule¹⁸, reacted with misfolded VHL in both immunofluorescence and immunoprecipitation experiments (Extended Data Fig. 5c–e). Notably, reactivity with the branched K11/K48 antibody required the presence of both K48- and K11-specific E3 ligases (Extended Data Fig. 5e). We conclude that a dual ubiquitin code involving both K11 and K48 is required for clearance of soluble PQC substrates. K11 ubiquitination is mediated by either of the membrane-bound E3s, Doa10 or Hrd1, and K48 ubiquitination is mediated by either of the soluble E3s, Ubr1 or San1, thus explaining the requirement for pairs of E3 ligases for PQC clearance. Given that deletion of one E3 pair did not abrogate the action of the other, addition of K11- or K48-linked chains does not require a particular sequential order.

It is unclear how chaperone proteins—known to be key mediators of PQC^{1,3,4}—facilitate ubiquitination and clearance. Chaperones could simply maintain the solubility of misfolded proteins, or they could specifically direct them along an E3 clearance pathway (Fig. 3a). We used the ubiquitin-linkage-specific ELISA assay to identify whether chaperones implicated in PQC are required to tag misfolded VHL with either K11-Ub or K48-Ub (Fig. 3b). We found that cells lacking Ssa1 and Ssa2—the major cytosolic proteins of the heat-shock protein (Hsp)70 family required for PQC clearance^{7,10,20}—were strongly impaired in ubiquitination with either linkage. By contrast, the

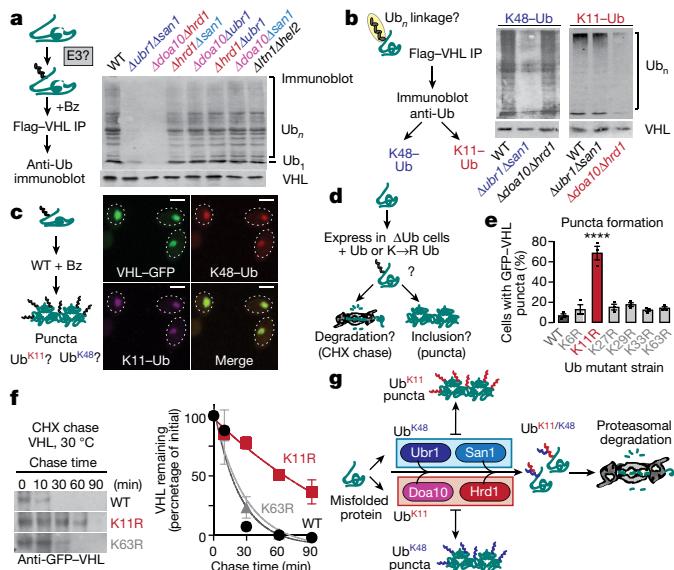


Fig. 2 | Cytoplasmic misfolded proteins are modified with both K11- and K48-linked ubiquitin chains. **a**, VHL ubiquitination is impaired in $\Delta\text{ubr1}\Delta\text{san1}$ and $\Delta\text{doa10}\Delta\text{hrd1}$ strains. Denaturing immunoprecipitation (IP) for Flag-VHL was followed by immunoblot for ubiquitin in WT or E3 double-deletion strains. Deletion of the co-translational E3s Ltn1 and Hel2 served as a control. **b**, K48-Ub and K11-Ub linkages are reduced on Flag-VHL in $\Delta\text{ubr1}\Delta\text{san1}$ and $\Delta\text{doa10}\Delta\text{hrd1}$ strains, respectively. Experiment performed as in panel **a**, but using K48-Ub or K11-Ub specific antibodies for immunoblot. **c**, K48-Ub and K11-Ub co-localize with GFP-VHL puncta. Experiment performed as in Fig. 1b. Fixed cells were spheroplasmed and immunostained before imaging by confocal fluorescence microscopy. Images represent more than 100 cells from each of three independent experiments. Scale bars represent 2 μm . **d–f**, VHL clearance is impaired in the absence of K11-Ub linkages. **d**, Cells co-expressing galactose-inducible GFP-VHL, with either WT or K-to-R mutant Ub as their only ubiquitin source, were shifted to glucose media for 1 h to shut off expression. 300 cells were counted per condition. **e**, Percentage of cells with GFP-VHL puncta (mean \pm s.e.m. from three biologically independent experiments). Only Ub^{K11R} significantly altered the percentage of puncta-positive cells versus WT (one-way ANOVA plus Dunnett's multiple comparisons test; $****P < 0.0001$). **f**, CHX chase and immunoblot to assess GFP-VHL stability in Ub mutant strains. Graphs represent densitometric quantification relative to $t = 0$ (mean \pm s.e.m. from three biologically independent experiments). **g**, Doa10/Hrd1 and Ubr1/San1 E3 ligases collaborate to ubiquitinate misfolded proteins with branched K11/K48 chains, thereby targeting them for proteasomal clearance. Inhibition of either type of linkage results in the sequestration of the misfolded proteins with only Ub^{K48} or Ub^{K11} into puncta. **a, b, f**, Immunoblots represent three biologically independent experiments.

ribosome-associated Hsp70s Ssb1 and Ssb2 were dispensable for degradation¹⁰ and ubiquitination.

Binding of Hsp70s to substrates relies on many J-domain proteins, themselves often chaperones that ferry substrates for Hsp70 binding^{3,4}. We examined VHL ubiquitination in cells lacking two J-domain proteins, Ydj1 and Sis1, that have been implicated in PQC^{7,9,11,20–22}. Strikingly, each J-domain protein reduced ubiquitination, but in a linkage-specific manner. We found that ΔYdj1 cells showed reduced K48 ubiquitination, albeit to a modest degree, probably because Ydj1 is partially redundant with other J-domain co-factors⁷. Depleting cells of the essential Sis1 left K48 ubiquitination unaffected but caused a dramatic loss of K11 ubiquitination. Of note, three other chaperones important for PQC—the Hsp70 nucleotide-exchange factor Sse1 (an Hsp110 chaperone), Hsp90, and Sti1/HOP (which bridges the interaction of Hsp70 with Hsp90)^{7,10,11,23}—were all required for K48 ubiquitination but were dispensable for K11 ubiquitination. We conclude that specific chaperone pathways direct PQC substrates to distinct E3 pathways to promote mixed linkage ubiquitination (Fig. 3c). Sis1 cooperates

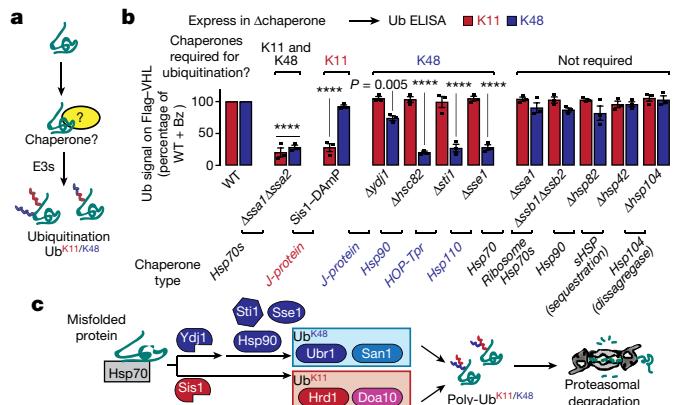


Fig. 3 | K11- and K48-linked ubiquitination of misfolded proteins involves different chaperones. **a**, Molecular chaperones are involved in the ubiquitination of misfolded proteins by E3 ligases. **b**, Relative amounts of Ub^{K11} and Ub^{K48} present on Flag-VHL in chaperone-deletion strains compared with WT. WT or chaperone-deletion strains expressing Flag-VHL were lysed after 1 h Bz treatment. ELISA for Ub linkages was then performed as described in Extended Data Fig. 3a. Bars represent normalized values from each strain (mean \pm s.e.m. from three biologically independent experiments) expressed as a proportion of normalized WT. Strains with statistically significant differences versus WT by one-way ANOVA plus Dunnett's multiple comparisons test are indicated with the adjusted P value or with $****$ for $P < 0.0001$. **c**, Ubiquitination of misfolded proteins through K11 and K48 linkages proceeds through the action of different chaperone pathways.

with Hsp70 in promoting Doa10/Hrd1-mediated K11 ubiquitination, whereas the Ydj1/Sse1/Sti1 chaperones cooperate with Hsp70 and Hsp90 for Ubr1/San1-mediated K48 ubiquitination. Requiring two distinct E3 ligases to communicate with distinct chaperone pathways may provide a checkpoint in the triage decision to refold or degrade.

That cytoplasmic misfolded proteins are degraded via a nuclear E3 ligase—San1—was puzzling. In principle, misfolded cytoplasmic proteins might be actively imported into the nucleus for degradation^{11,12,20,23}. Alternatively, they might passively diffuse through the nuclear pores, owing to their small size²⁴. To directly investigate cytoplasmic-specific and nuclear-specific PQC degradation pathways, we spatially restricted two PQC substrates—VHL and luciferase^{ts}—to the nucleus by using a nuclear localization signal (NLS), or to the cytoplasm by using a nuclear export signal (NES; Fig. 4a). Similar results were obtained for both substrates, corroborating the generality of our conclusions. Treatment with bortezomib showed that both nuclear and cytoplasmic variants are degraded by the UPS (Fig. 4b, c and Extended Data Fig. 6a). NES-GFP-VHL accumulated in cytoplasmic perinuclear puncta, whereas NLS-GFP-VHL accumulated in intranuclear puncta (Fig. 4b). Thus, misfolded proteins are either degraded or form inclusions in the cellular compartment where misfolding occurs.

Degradation of cytoplasmic NES-VHL or NES-luciferase^{ts} required the K11 ubiquitin ligases Doa10 and Hrd1, but only the cytoplasmic K48 ligase Ubr1 (Fig. 4d and Extended Data Fig. 6b). Therefore, San1 is dispensable for UPS degradation of strictly cytoplasmic PQC substrates. Surprisingly, the nuclear NLS-GFP-VHL or luciferase^{ts} required only the nuclear K48 ligase, San1, for clearance.

Consistent with their E3 requirements, cytoplasm-restricted misfolded NES-labelled proteins were conjugated to both K11- and K48-linked chains at levels similar to those of their unmodified counterparts. However, nuclear-restricted misfolded NLS-labelled proteins were ubiquitinated only with K48 chains, with the K11 signal reduced to baseline levels (Fig. 4e and Extended Data Fig. 6c). Notably, nuclear misfolded proteins tagged with K48-Ub were efficiently cleared by the proteasome (Fig. 4c, d and Extended Data Fig. 6a). Confirming the distinct role of K11 ubiquitination in nuclear versus cytoplasmic PQC, clearance of NES-tagged misfolded proteins was impaired in Ub^{K11R}

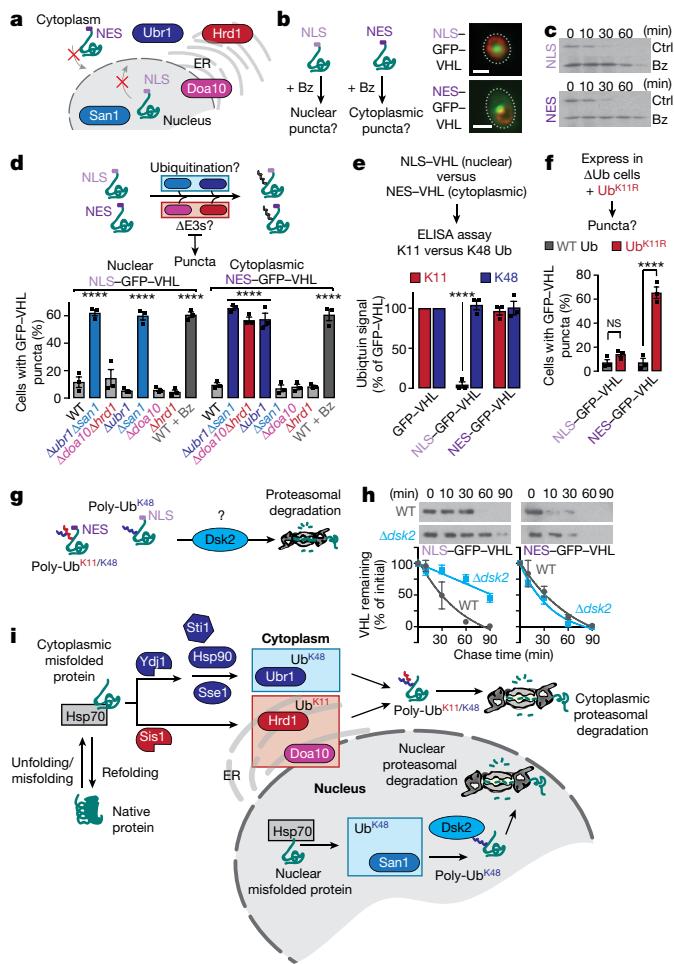


Fig. 4 | Confining misfolded proteins to the nucleus or cytoplasm alters their PQC requirements. **a, b**, Upon proteasome inhibition, NLS-GFP-VHL (NLS here) and NES-GFP-VHL (NES) accumulate in the nucleus or cytoplasm, respectively. (The red crosses indicate that the proteins are not transported across the nuclear envelope, because of the presence of the NLS or NES.) The locations of the E3s Ubr1, Hrd1, Doa10 and San1 are also shown. **b**, Expression of NLS-GFP-VHL or NES-GFP-VHL in WT cells was shut off in glucose media with 50 μ M Bz for 1 h. Cells were immunostained for Nsp1 (a nuclear pore protein; red) and imaged by fluorescence microscopy. Images represent three biologically independent experiments. Scale bars represent 2 μ m. **c**, NLS-GFP-VHL and NES-GFP-VHL are cleared by the proteasome. Cycloheximide chase and immunoblot were used to assess the stability of these proteins in WT cells treated with (Bz) or without (Ctrl) 50 μ M Bz. Immunoblots represent three biologically independent experiments. **d**, Confining VHL to the nucleus or cytoplasm alters its E3 requirement. The graph shows the percentage of cells with NLS-GFP-VHL or NES-GFP-VHL puncta in deletion strains following 1 h of shut-off. **e**, Nuclear VHL has severely reduced K11-Ub ubiquitin linkages. ELISA performed as in Fig. 2c, but using GFP-multiTrap instead of α -Flag-conjugated plates. **f**, Nuclear VHL clearance is unaffected by K11-Ub linkages. Experiment performed as in panel **d**, but using cells expressing WT ubiquitin or Ub^{K11R} as the only ubiquitin source. **g, h**, Clearance of NLS-GFP-VHL, but not NES-GFP-VHL, requires Dsk2. CHX chase performed as in **c**, but in WT or Δ dsk2 cells. The densitometric quantification in panel **h** is shown relative to $t=0$ (mean \pm s.e.m. from three biologically independent experiments). **i**, Nuclear and cytoplasmic misfolded proteins have distinct clearance requirements. Cytoplasmic misfolded proteins require tagging with both K11-Ub and K48-Ub by chaperones and E3 ligases for proteasomal degradation. In the nucleus, tagging with K48-Ub is sufficient for recognition by Dsk2 and subsequent proteasomal degradation. **d–f**, Bars represent mean \pm s.e.m. from three biologically independent experiments. Statistically significant differences versus WT (**d, f**) or GFP-VHL (**e**) by one-way ANOVA plus Dunnett's multiple comparisons test (**d, e**), or two-tailed Student's *t*-test (**f**), are indicated (adjusted *P* value, or $****P < 0.0001$; NS, *P* > 0.05).

cells, whereas degradation of nuclear-restricted PQC substrates was unaffected (Fig. 4f and Extended Data 6d).

To understand the differences between nuclear and cytoplasmic PQC, we first focused on the ubiquilin Dsk2—a predominantly nuclear²⁵ shuttling factor that physically ferries K48-ubiquitinated proteins to the proteasome²⁶. Strikingly, cells lacking Dsk2 were impaired in clearance of NLS-tagged but not NES-tagged GFP-VHL (Fig. 4g, h). We propose that Dsk2 in the nucleus increases the affinity of K48-only nuclear misfolded proteins for the proteasome.

We examined the role of chaperone circuits in nuclear versus cytoplasmic PQC through functional and proteomic approaches. We found that depletion of Sis1—required for K11 ubiquitination (Fig. 3b)—had no effect on the clearance of NLS-tagged proteins, but, as expected, blocked the clearance of NES-tagged variants (Extended Data Fig. 6d). Interestingly, Hsp90, Sti1, Sse1 and Ydj1 were required only for K48 ubiquitination and clearance of cytoplasmic substrates. Only the Hsp70 proteins—Ssa1 and Ssa2—were required for nuclear PQC.

Mass spectrometry based on stable-isotope labelling by amino acids in cell culture (SILAC) further indicated that PQC of nuclear and cytoplasmic misfolded proteins involves different circuitries. Analysis of immunoprecipitates of NLS-tagged or NES-tagged GFP-VHL (Extended Data Fig. 7 and Extended Table 1) revealed localization-dependent proteostasis interactors (with a $\log_2(VHL/\text{control ratio})$ of greater than 0.5). Both nuclear and cytoplasmic PQC substrates shared enrichments in proteasomal subunits and several chaperones, including four Hsp70 proteins. Consistent with their selective requirement for cytoplasmic PQC, Hsp82 and Sis1 specifically associated with NES-tagged VHL. Of note, Cdc48/p97 also selectively associated with NES-tagged VHL, whereas the TRiC/CCT chaperonin selectively associated with NLS-tagged VHL (Extended Data Fig. 7c). The importance of these interactions for PQC will be explored in future studies.

We conclude that the clearance of nuclear and cytoplasmic PQC substrates requires distinct ubiquitin codes, distinct E3 ligases, and distinct chaperone sets. Cytoplasmic PQC requires both K11- and K48-linked ubiquitin, while nuclear PQC requires only K48-linked ubiquitin. Given that the misfolded protein is the same in both compartments, these distinct requirements are unlikely to relate to the substrate's structural properties, but instead arise from differences in the proteostasis machineries that maintain the nuclear and cytoplasmic proteomes.

Our study opens new perspectives for understanding the circuits and logic of misfolded-protein quality control (Fig. 4i). We define a general PQC path through which specific, non-redundant networks of E3 ligases and chaperones mediate post-translational PQC clearance. Surprisingly, we also uncover marked differences in the pathways of cytoplasmic and nuclear PQC. It is possible that nuclear and cytoplasmic proteasomes differ in composition²⁷ or concentration²⁸, affecting their ability to recognize ubiquitinated substrates effectively. It is also possible that K11-linked ubiquitin chains facilitate recognition by another cytoplasmic PQC factor. Intriguingly, Sis1, required for K11 linkages, can act as a sorting factor⁹, and is sequestered in protein aggregates^{7–9,21,29}. Nuclear ubiquilin Dsk2 shuttles K48-linked substrates to nuclear proteasomes, whereas in the cytosol, mixed K11/K48 chains enhance the affinity of misfolded proteins for cytoplasmic proteasomes, probably by engaging multiple ubiquitin receptors in the 19S proteasome cap³⁰.

The distinct ubiquitin linkage requirements for nuclear and cytoplasmic PQC might respond to distinct regulatory and functional rationales for the triage decision between (re-)folding and targeting for degradation. For example, the importance of the UPS in chromatin regulation³¹ could sensitize the nuclear proteome to protein misfolding and aggregation³²—perhaps leading to relaxed ubiquitination requirements for proteasomal recognition. By contrast, the cytosol is a site of active protein biogenesis, assembly and targeting, requiring productive folding intermediates to be safeguarded. Requiring a dual ubiquitin code for cytoplasmic PQC—mediated by distinct E3 ligases and chaperones—would provide a checkpoint to ensure that only non-productive, misfolded proteins are degraded. Of interest, ubiquilin and the PQC

E3 ligases identified here are associated with a host of human diseases (Extended Table 2). Dissection of these circuits in normal and diseased states might provide mechanistic clues and open up therapeutic opportunities.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0678-x>.

Received: 18 December 2017; Accepted: 4 September 2018;

Published online 31 October 2018.

1. Chiti, F. & Dobson, C. M. Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. *Annu. Rev. Biochem.* **86**, 27–68 (2017).
2. Balch, W. E., Morimoto, R. I., Dillin, A. & Kelly, J. W. Adapting proteostasis for disease intervention. *Science* **319**, 916–919 (2008).
3. Sontag, E. M., Samant, R. S. & Frydman, J. Mechanisms and functions of spatial protein quality control. *Annu. Rev. Biochem.* **86**, 97–122 (2017).
4. Balchin, D., Hayer-Hartl, M. & Hartl, F. U. *In vivo* aspects of protein folding and quality control. *Science* **353**, aac4354 (2016).
5. Kwon, Y. T. & Ciechanover, A. The ubiquitin code in the ubiquitin-proteasome system and autophagy. *Trends Biochem. Sci.* **42**, 873–886 (2017).
6. Kaganovich, D., Kopito, R. & Frydman, J. Misfolded proteins partition between two distinct quality control compartments. *Nature* **454**, 1088–1095 (2008).
7. Escusa-Toret, S., Vonk, W. I. & Frydman, J. Spatial sequestration of misfolded proteins by a dynamic chaperone pathway enhances cellular fitness during stress. *Nat. Cell Biol.* **15**, 1231–1243 (2013).
8. Malinovska, L., Kroschwald, S., Munder, M. C., Richter, D. & Alberti, S. Molecular chaperones and stress-inducible protein-sorting factors coordinate the spatiotemporal distribution of protein aggregates. *Mol. Biol. Cell* **23**, 3041–3056 (2012).
9. Park, S. H. et al. PolyQ proteins interfere with nuclear degradation of cytosolic proteins by sequestering the Sis1p chaperone. *Cell* **154**, 134–145 (2013).
10. McClellan, A. J., Scott, M. D. & Frydman, J. Folding and quality control of the VHL tumor suppressor proceed through distinct chaperone pathways. *Cell* **121**, 739–748 (2005).
11. Heck, J. W., Cheung, S. K. & Hampton, R. Y. Cytoplasmic protein quality control degradation mediated by parallel actions of the E3 ubiquitin ligases Ubr1 and San1. *Proc. Natl. Acad. Sci. USA* **107**, 1106–1111 (2010).
12. Prasad, R., Kawaguchi, S. & Ng, D. T. A nucleus-based quality control mechanism for cytosolic proteins. *Mol. Biol. Cell* **21**, 2117–2127 (2010).
13. Deng, M. & Hochstrasser, M. Spatially regulated ubiquitin ligation by an ER/nuclear membrane ligase. *Nature* **443**, 827–831 (2006).
14. Swanson, R., Locher, M. & Hochstrasser, M. A conserved ubiquitin ligase of the nuclear envelope/endoplasmic reticulum that functions in both ER-associated and Matalpha2 repressor degradation. *Genes Dev.* **15**, 2660–2674 (2001).
15. Jin, L., Williamson, A., Banerjee, S., Philipp, I. & Rape, M. Mechanism of ubiquitin-chain formation by the human anaphase-promoting complex. *Cell* **133**, 653–665 (2008).
16. Xu, P. et al. Quantitative proteomics reveals the function of unconventional ubiquitin chains in proteasomal degradation. *Cell* **137**, 133–145 (2009).
17. Yau, R. & Rape, M. The increasing complexity of the ubiquitin code. *Nat. Cell Biol.* **18**, 579–586 (2016).
18. Yau, R. G. et al. Assembly and function of heterotypic ubiquitin chains in cell-cycle and protein quality control. *Cell* **171**, 918–933 (2017).
19. Spence, J., Sadis, S., Haas, A. L. & Finley, D. A ubiquitin mutant with specific defects in DNA repair and multiubiquitination. *Mol. Cell. Biol.* **15**, 1265–1273 (1995).
20. Prasad, R., Xu, C. & Ng, D. T. W. Hsp40/70/110 chaperones adapt nuclear protein quality control to serve cytosolic clients. *J. Cell Biol.* **217**, 2019–2032 (2018).
21. Summers, D. W., Wolfe, K. J., Ren, H. Y. & Cyr, D. M. The type II Hsp40 Sis1 cooperates with Hsp70 and the E3 ligase Ubr1 to promote degradation of terminally misfolded cytosolic protein. *PLoS One* **8**, e52099 (2013).
22. Shiber, A., Breuer, W., Brandeis, M. & Ravid, T. Ubiquitin conjugation triggers misfolded protein sequestration into quality control foci when Hsp70 chaperone levels are limiting. *Mol. Biol. Cell* **24**, 2076–2087 (2013).
23. Guerrero, C. J., Weiberth, K. F. & Brodsky, J. L. Hsp70 targets a cytoplasmic quality control substrate to the San1p ubiquitin ligase. *J. Biol. Chem.* **288**, 18506–18520 (2013).
24. Amm, I. & Wolf, D. H. Molecular mass as a determinant for nuclear San1-dependent targeting of misfolded cytosolic proteins to proteasomal degradation. *FEBS Lett.* **590**, 1765–1775 (2016).
25. Biggins, S., Ivanovska, I. & Rose, M. D. Yeast ubiquitin-like genes are involved in duplication of the microtubule organizing center. *J. Cell Biol.* **133**, 1331–1346 (1996).
26. Tsuchiya, H. et al. *In vivo* ubiquitin linkage-type analysis reveals that the Cdc48-Rad23/Dsk2 axis contributes to K48-linked chain specificity of the proteasome. *Mol. Cell* **66**, 488–502 (2017).
27. Fabre, B. et al. Subcellular distribution and dynamics of active proteasome complexes unraveled by a workflow combining *in vivo* complex cross-linking and quantitative proteomics. *Mol. Cell. Proteomics* **12**, 687–699 (2013).
28. Russell, S. J., Steger, K. A. & Johnston, S. A. Subcellular localization, stoichiometry, and protein levels of 26 S proteasome subunits in yeast. *J. Biol. Chem.* **274**, 21943–21952 (1999).
29. Miller, S. B. et al. Compartment-specific aggregates direct distinct nuclear and cytoplasmic aggregate deposition. *EMBO J.* **34**, 778–797 (2015).
30. Chen, X. et al. Structures of Rpn1 T1:Rad23 and hRpn13:hPLIC2 reveal distinct binding mechanisms between substrate receptors and shuttle factors of the proteasome. *Structure* **24**, 1257–1270 (2016).
31. Ben Yehuda, A. et al. Ubiquitin accumulation on disease associated protein aggregates is correlated with nuclear ubiquitin depletion, histone de-ubiquitination and impaired DNA damage response. *PLoS One* **12**, e0169054 (2017).
32. Zhong, Y. et al. Nuclear export of misfolded SOD1 mediated by a normally buried NES-like sequence reduces proteotoxicity in the nucleus. *eLife* **6**, e23759 (2017).

Acknowledgements We thank K. Li, M. Burlingame and A. L. Burlingame for help with mass spectrometry; R. Andino and F. U. Hartl for critical reading of the manuscript; D. R. Gestaut for sharing the NLS- and NES-tagged plasmids; and all members of the Frydman laboratory for advice. R.S.S. was supported by a Human Frontier Science Program long-term fellowship (LT000695/2015-L). C.M.L. was supported by a National Institutes of Health (NIH) postdoctoral fellowship (1F32CA162919-01A1). This work was supported by an NIH grant (R37GM056433) to J.F. E.M.S. was supported by a postdoctoral fellowship from NIH (F32NS086253).

Reviewer information *Nature* thanks I. Dikic, A. Dillin and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions R.S.S. and J.F. designed the study. C.M.L. performed the initial puncta screens with E3 single- and double-deletion mutants. R.S.S. performed all other experiments and analysis. E.M.S. provided insight into the contribution of Dsk2 to nuclear quality control. R.S.S. and J.F. interpreted the data and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0678-x>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0678-x>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to J.F.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Yeast media, plasmids and strains. Preparation of yeast media, growth, transformations and manipulations were performed according to standard protocols. All E3- and chaperone-deletion yeast strains were derived from the BY4742 wild-type strain. Single deletions were generated by homologous recombination using the *NAT* gene. The *Sis1*–DAmP strain was also generated in this way. Double deletions were generated with both *NAT* and hygromycin. All strains were checked by polymerase chain reaction (PCR) using at least two sets of primers.

All ubiquitin K-to-R mutant strains—expressing a single, galactose-inducible ubiquitin gene—were gifts from D. Finley (Harvard Univ., MA, USA)²³. Yeast strains expressing GFP-tagged Doa10 or Hrd1 from their endogenous loci were from the Yeast-GFP Clone Collection (Thermo Fisher Scientific). The *Δdsk2* strain was obtained from the *Saccharomyces* Genome Deletion Project³³. We acknowledge gifts of pGAL-CPY⁺-GFP (R. Hampton, Univ. California San Diego, CA, USA), pADH-Flag-Ubr1 and pGAL-San1-V5His6 (D. Wolf, Univ. Stuttgart, Germany), pFLUC-DM-YFP (F. U. Hartl, Martinsried, Germany), and Ub-M-GFP, Ub-R-GFP, Ub^{G76V}-GFP and GFP-CL1 (N. Dantuma, Karolinska Institute, Stockholm, Sweden)³⁴. All other plasmids were constructed using the Gateway cloning technology³⁵ as described².

Galactose shut-off protein expression. Yeast strains transformed with plasmids encoding the galactose-inducible protein of interest were grown overnight in raffinose synthetic medium at 30 °C before dilution to an optical density at 600 nm (OD₆₀₀) of between 0.05 and 0.1 in galactose synthetic medium. The cells were grown for 4–6 h (OD₆₀₀ 0.6–0.8) to induce expression of the galactose-inducible protein. Expression was shut off by switching the cells to glucose synthetic medium, and the fate of the existing pool of proteins was assessed according to the appropriate downstream application.

Counting puncta-containing cells. Cells were grown as described for galactose shut-off protein expression. Following shut-off, cells were allowed to grow at 30 °C or 37 °C for 1 h in glucose synthetic medium. Note that for WT plus bortezomib conditions, 50 μM bortezomib (LC Laboratories) was dissolved in the glucose synthetic medium before addition. Cells were then fixed for 15 min in 4% paraformaldehyde before mounting onto concanavalin-A-coated coverslips using ProLongTM Diamond Antifade Mountant with DAPI (Thermo Fisher Scientific). Fluorescence was visualized using a Zeiss LSM700 confocal microscope with a ×63 oil-immersion lens. Image analysis was performed by ImageJ software (<http://imagej.nih.gov/ij>). 300 cells were counted manually for each blinded sample, and the percentage of cells containing GFP-positive puncta was noted. Statistical analysis was performed using one-way ANOVA followed by Dunnett's multiple comparisons test.

Immunofluorescence. Cells were grown and paraformaldehyde fixed as described above. Fixed cells were spheroplasted by incubating for 20–40 min at 30 °C with Zymolyase 100T (Zymo Research) in potassium phosphate buffer (0.1 M potassium phosphate pH 7.5, 1.2 M sorbitol) supplemented with 25 mM DTT and 5 mM EDTA. The resultant spheroplasts were permeabilized with 0.1% (v/v) Triton X-100 in potassium phosphate buffer for 10 min. For immunostaining, cells were blocked (potassium phosphate buffer with 1% (w/v) bovine serum albumin (BSA) for 30 min at room temperature) and incubated overnight at 4 °C with antibodies diluted at the appropriate concentration in potassium phosphate buffer with 0.1% (w/v) BSA. Antibodies used were against K48-Ub covalently linked to Alexa Fluor 568 (1/100; Abcam catalogue number ab208136), K11-Ub (1/50; EMD Millipore catalogue number MABS107-I) covalently linked to Alexa Fluor 647 NHS Ester (Thermo Fisher Scientific), K11/K48-Ub bispecific antibody (1/500; Genentech) with secondary antibody Cy5-conjugated donkey anti-human (1/1,000; Jackson ImmunoResearch catalogue number 709-175-149), and Nsp1 (1:500, EnCor catalogue number MCA-32D6) with secondary antibody Alexa Fluor 568-conjugated goat anti-mouse (1/1,000; Thermo Fisher Scientific catalogue number A10037) for 1 h at room temperature. Stained cells were mounted onto polylysine-coated coverslips using ProLong Diamond Antifade Mountant with DAPI (Thermo Fisher Scientific). Fluorescence was visualized using a Zeiss LSM700 confocal microscope with ×100, numerical aperture 1.4, oil-immersion lens. Raw data collected as z-stacks were represented in a single image as maximum-intensity projections (ImageJ).

Cycloheximide chase assay. Cells were grown as described for galactose shut-off protein expression. A first sample with an equivalent OD₆₀₀ of 10 (*t* = 0) was collected before shut-off. The rest of the culture was shifted to glucose synthetic medium with 50 μg ml⁻¹ cycloheximide, and further cells (OD₆₀₀ = 10) were collected 10, 30, 60 and 90 min after the shift. Samples from each time point were pelleted, washed once with 15 mM sodium azide containing 1× Roche cCompleteTM EDTA-free protease-inhibitor tablet (Sigma), snap frozen in liquid nitrogen, and stored at -80 °C until all time points were collected. Proteins were extracted by boiling each sample in an equal volume of 2× SDS sample buffer (100 mM Tris-HCl pH 6.8, 4% (w/v) SDS, 20% (v/v) glycerol, 200 mM DTT, 0.2% (w/v) bromophenol blue) for 10 min before detection of protein by SDS-PAGE and immunoblotting.

SDS-PAGE and immunoblotting. Protein samples from cell lysates or immunoprecipitates were denatured in SDS sample buffer (95 °C for 5 min) or LDS sample buffer (70 °C for 10 min) before separation by SDS-PAGE. Precision Plus prestained protein standards (Bio-Rad) were used to estimate protein weight. Proteins were transferred onto polyvinylidene fluoride (PVDF) or nitrocellulose membranes (Bio-Rad) and immunoblotted with primary antibodies against GFP (1/1,000; Roche catalogue number 11814460001 or Santa Cruz Biotechnology catalogue number sc-9996), peroxidase anti-peroxidase complex for detection of Protein A in the tandem affinity purification tag (TAP; 1/2,000 to 1/7,500; Sigma catalogue number P1291), glyceraldehyde 3-phosphate dehydrogenase (GAPDH; 1/5,000, Abcam catalogue number ab9485; or 1/10,000, Genetex catalogue number GTX100118), α-tubulin (1/2,500; DSHB Hybridoma Product 12G10; deposited by J. Frankel and E.M. Nelson), Flag (1/1,000; Cell Signaling Technology catalogue number 2368), pan-ubiquitin (1/1,000; Life Sensors catalogue number VU101), K11-Ub (1/100; EMD Millipore catalogue number MABS107-I), K48-Ub (1/1,000; Cell Signaling Technology catalogue number 12805) or K11/K48-Ub bispecific antibody (1/500; Genentech). Specific primary antibodies used are indicated next to the uncropped immunoblot images in Supplementary Fig. 1. For immunoblotting of ubiquitin, samples were separated by SDS-PAGE, transferred to a PVDF membrane and denatured by boiling for 10 min at 95 °C before antibody incubation. Secondary antibodies used were horseradish peroxidase (HRP)-conjugated donkey anti-mouse (1/5,000; Jackson ImmunoResearch catalogue number 715-035-150), donkey anti-rabbit (1/5,000; Jackson ImmunoResearch catalogue number 711-035-152) or donkey anti-human (1/5,000; Jackson ImmunoResearch catalogue number 709-035-149). The HRP signal was detected by incubation with Pierce ECL Western blotting substrate (Thermo Fisher Scientific) and exposure to GeneMate Blue Ultra Film (BioExpress). Immunoblots shown are representative of three independent experiments.

Immunoprecipitation of E3 ligases. To test for co-immunoprecipitation of E3 ligases, we transfected pADH-Flag-Ubr1 or pGAL-San1-V5His6 plasmids into yeast strains from the yeast-GFP collection (expressing GFP-tagged Doa10 or Hrd1 from the endogenous loci). Each strain was grown overnight in raffinose synthetic media at 30 °C before dilution to an OD₆₀₀ of between 0.05 and 0.1 in galactose synthetic media. The cells were grown for 24 h to induce expression of the galactose-inducible San1-V5His6, diluted back to OD₆₀₀ = 0.1, and then grown for another 4–6 h (OD₆₀₀ 0.6–0.8). For consistency between experimental conditions, the same protocol was followed for cells expressing Flag-Ubr1 from the alcohol dehydrogenase (ADH) promoter, even though this does not require galactose for expression. Cells were pelleted and washed once with 15 mM sodium azide supplemented with 1× Roche cCompleteTM EDTA-free protease-inhibitor tablet (Sigma) and 50 mM 2-chloroacetamide, to inhibit proteases and deubiquitinases, respectively. Pellets were resuspended in an equal volume of lysis buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 2 mM EDTA, 1 mM phenylmethylene sulfonyl fluoride (PMSF) Roche cCompleteTM EDTA-free protease-inhibitor tablet (Sigma), 50 mM 2-chloroacetamide, 10 μM PR-619 (Sigma)) and frozen drop-wise in liquid nitrogen by passing through a 20.5-gauge syringe. Frozen samples were lysed by cryogrinding in a Retsch MM-301 (five cycles, 30 Hz, for 3 min per cycle) and proteins solubilized by adding Triton X-100 (1% v/v final concentration). Lysates were clarified (16,000g for 30 min at 4 °C) and quantified for total protein by bicinchoninic acid (BCA) assay. We incubated 2 mg of lysate with anti-GFP rabbit IgG conjugated to protein G dynabeads (Thermo Fisher Scientific) for 2 h at 4 °C to immunoprecipitate the GFP-tagged protein complexes, which were then eluted from the beads by heating for 30 min at 70 °C in non-reducing LDS sample buffer (Thermo Fisher Scientific). The bead-free samples were reduced with DTT (50 mM final concentration, 10 min at 70 °C) before SDS-PAGE analysis.

Immunoprecipitation of ubiquitinated VHL. For immunoprecipitation of ubiquitinated Flag-VHL, yeast strains were grown as described for galactose shut-off protein expression. Following shut-off in glucose synthetic medium supplemented with 50 μM Bz for 1 h at 30 °C, cells were pelleted and snap frozen in liquid nitrogen. All subsequent steps were performed at 4 °C or on ice. Pellets were resuspended in an equal volume of urea lysis buffer (50 mM Tris-HCl pH 7.5, 8 M urea, 150 mM NaCl, 2 mM EDTA, 1 mM PMSF, Roche cCompleteTM EDTA-free protease inhibitor tablet (Sigma), 50 mM chloroacetamide, 10 μM PR-619 (Sigma)) and lysed by bead beating (five cycles at 1 min each, with 1 min on ice in between cycles). Following dilution tenfold in Triton immunoprecipitation buffer (same composition as urea lysis buffer, but with 1% v/v Triton X-100 instead of 8 M urea), lysates were clarified (16,000g for 30 min at 4 °C) and quantified for total protein by BCA assay. We incubated 2 mg of lysates with Flag-M2 magnetic beads (Sigma) for 2 h at 4 °C to immunoprecipitate the Flag-tagged protein, which was then eluted from the beads by heating for 30 min at 70 °C in non-reducing LDS sample buffer (Thermo Fisher Scientific), to avoid co-elution of the Flag antibody. The bead-free samples were reduced with DTT (50 mM final concentration, 10 min at 70 °C) before SDS-PAGE analysis.

The same protocol was used for denaturing immunoprecipitation of ubiquitinated GFP-VHL, but with addition of 1% w/v SDS in lysis and immunoprecipitation

buffers, and incubation with GFP-TRAP_MA magnetic beads (ChromoTek) instead of Flag-M2 magnetic beads.

For double-immunoprecipitation experiments, 10 mg of cell lysate was incubated with Flag-M2 magnetic beads for 2 h at 4°C and eluted from the beads by competition with 3× Flag peptide (Apex Bio). The resultant eluate was subsequently incubated with an antibody against K48-Ub (1/500; Cell Signaling Technologies catalogue number 4289) or K11-Ub (1/50; EMD Millipore catalogue number MABS107-I) covalently conjugated using bis(sulfosuccinimidyl) suberate (BS³) to protein G dynabeads (Thermo Fisher Scientific) overnight at 4°C. The bead-bound fraction ('eluate') was eluted by heating for 10 min at 70°C in non-reducing LDS sample buffer, and analysed alongside the unbound fraction ('flow-through') by SDS-PAGE.

Ubiquitin linkage ELISA. For quantification of K11-Ub and K48-Ub linkages on Flag-VHL, cell lysates were prepared as described for immunoprecipitation of ubiquitinated Flag-VHL. We added 200 µg of lysate protein to each well of an anti-Flag-M2-coated 96-well plate and then incubated the plate for 2 h. All incubation steps were performed at room temperature with gentle shaking. Four wells were used for each technical replicate (two replicates per strain per experiment). After washing four times with Triton immunoprecipitation buffer to remove unbound protein, each well was incubated for 1 h with rabbit antibodies against one of GFP (1/1,000; Cell Signaling Technology catalogue number 2956), Flag (1/1,000; Cell Signaling Technology catalogue number 2368), K11-Ub (1/50; EMD Millipore catalogue number MABS107-I), or K48-Ub (1/500; Cell Signaling Technology catalogue number 4289) diluted in 100 µl tris-buffered saline/Tween-20 (TBS-T) buffer with 0.1% BSA. After another four washes, each well was incubated for 1 h with HRP-conjugated donkey anti-rabbit antibody (1/2,000; Jackson ImmunoResearch catalogue number 711-035-152) diluted in 100 µl TBS-T with 0.1% BSA, washed another four times with Tween immunoprecipitation buffer, and incubated for 30 min with 100 µl Pierce tetramethylbenzidine (TMB) substrate (Thermo Fisher Scientific) followed by 100 µl 0.16 M sulfuric acid to stop the reaction. Absorbance was measured at 450 nm.

To calculate the K11-Ub or K48-Ub signal for each strain, we subtracted the raw absorbance readings of the negative control (GFP) signal, and then divided by the Flag signal to account for variations in total Flag-VHL levels. These K11-Ub or K48-Ub signals were then expressed as a proportion of the K11-Ub or K48-Ub signal in the WT strain to allow direct comparison between strains. Bars represent means ± s.e.m. from three individual experiments.

The same protocol was followed for quantification of linkages on GFP-tagged proteins, except with the use of GFP-multiTrap 96-well plates (ChromoTek) instead of Flag-M2-coated plates, and using the GFP and Flag signals as positive and negative controls, respectively.

SILAC mass spectrometry of VHL immunoprecipitates. WT yeast cells transfected with one of NLS-GFP-VHL, NES-GFP-VHL or Flag-VHL were grown overnight at 30°C in raffinose-synthetic media supplemented with light Lys0 (Cambridge Isotope Laboratories catalogue number ULM-8766-PK), heavy Lys8 (Cambridge Isotope Laboratories catalogue number CNLM-291-H-1) or medium Lys4 (Cambridge Isotope Laboratories catalogue number DLM-2640-PK), respectively. Cells were diluted to an OD₆₀₀ of between 0.05 and 0.1 in galactose synthetic medium supplemented with the appropriate lysine isotopes and grown for 4–5 h (OD₆₀₀ 0.6–0.8) to induce expression of the galactose-inducible protein. Expression was shut off by switching the cells to glucose synthetic medium supplemented with the appropriate lysine isotope. Pelleted cells were lysed by cryogrinding as described in the 'Immunoprecipitation of E3 ligases' section above. Then, 1.5 mg of protein (as quantified by BCA assay) from each of the NLS-GFP-VHL, NES-GFP-VHL and Flag-VHL lysates were mixed before immunoprecipitation using GFP-TRAP_MA magnetic beads (ChromoTek) according to the manufacturer's protocol. After three washes with Triton immunoprecipitation buffer, the beads were washed twice with 50 mM Tris-HCl pH 8 supplemented with 20 mM CaCl₂. On-bead trypsin digestion and peptide clean-up were performed using the in-StageTip method³⁶. Peptides were analysed on a Q Exactive Plus Orbitrap (Thermo Fisher Scientific) connected to a NanoAcuity high-performance liquid-chromatography system (Waters). An EASY-Spray PepMap rapid-separation liquid-chromatography column (C18, 3 µm, 75 µm × 15 cm; Thermo Fisher Scientific) was used to resolve peptides with a binary solvent system (0.1% formic acid in water as mobile phase A, 0.1% formic acid in acetonitrile as mobile phase B). The Q Exactive Plus was run on a linear 60-min gradient from 2% to 30% phase B at a flow rate of 300 nL min⁻¹. Both precursor and fragment ions were analysed in the flow-through (FT) mode at a mass resolution of 70,000 and 17,500, respectively. After a survey scan, the ten most intense precursor ions were selected for subsequent fragmentation by higher-energy collisional dissociation.

Raw data from four biological replicates were processed using MaxQuant³⁷ (<http://www.coxdocs.org/doku.php?id=maxquant:start>, version 1.6.2.3) and searched against the *Saccharomyces* Genome Database (https://downloads.yeast-genome.org/sequence/S288C_reference/orf_protein; downloaded in January

2015) with common contaminant entries. The default MaxQuant parameters for a triple SILAC experiment were used, with the exception of 'Re-quantify', which was enabled.

The proteinGroups.txt file was filtered to exclude contaminants, reverse hits, hits 'only identified by site', and hits for which only one peptide was identified. The normalized SILAC ratios were used to generate median fold-change values per protein. Proteins with a log₂(light/medium) or log₂(heavy/medium) value of more than 0.5 were counted as 'enriched' in NLS-VHL or NES-VHL interactomes, respectively. Enriched proteins from each interactome were subjected to pathway analysis to search for enriched Gene Ontology (GO) terms, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and PFAM protein domains in either interactome using the STRING database³⁸ (<http://string-db.org>, version 10.5).

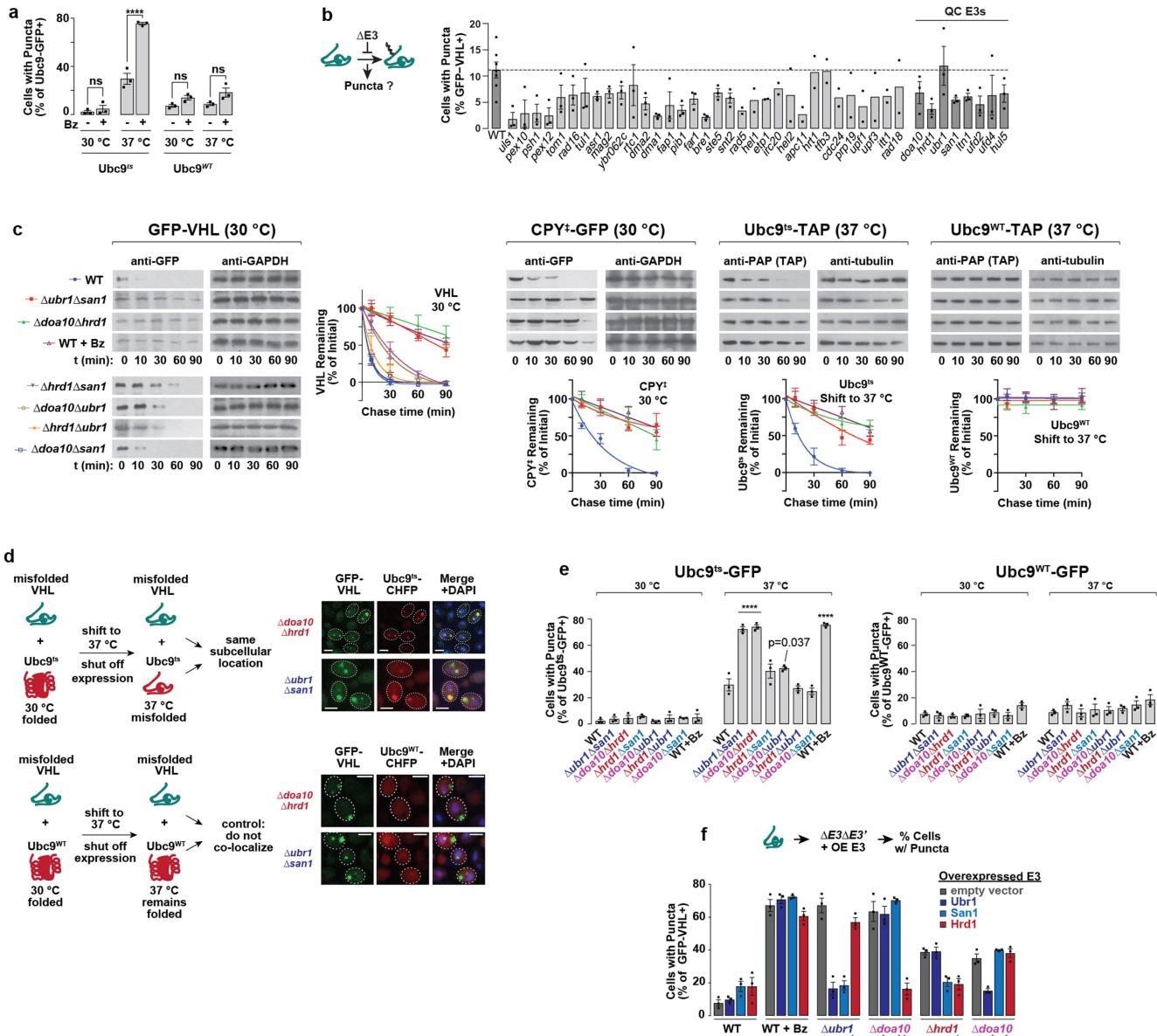
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

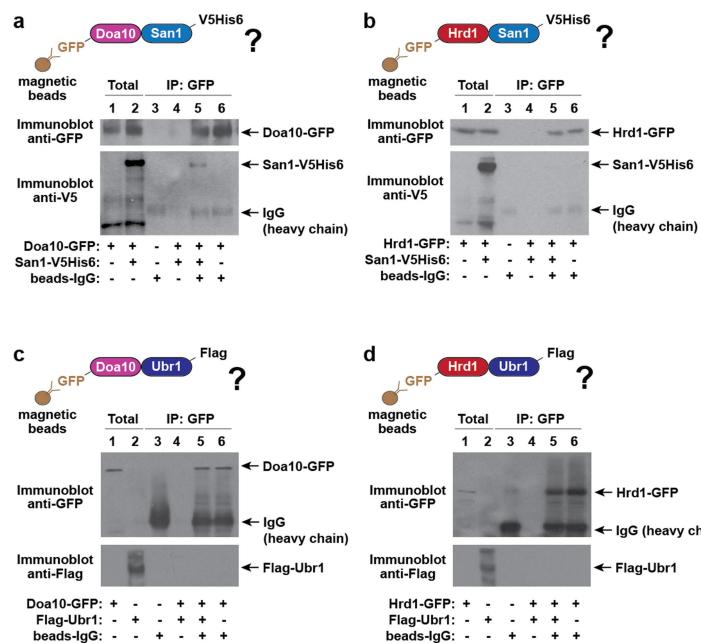
The data sets generated and/or analysed during this study are available from the corresponding author on reasonable request. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner repository with the data identifier PXD010660. Uncropped images of all immunoblots shown in this study are in Supplementary Fig. 1.

33. Winzeler, E. A. et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
34. Dantuma, N. P., Lindsten, K., Glas, R., Jellne, M. & Masucci, M. G. Short-lived green fluorescent proteins for quantifying ubiquitin/proteasome-dependent proteolysis in living cells. *Nat. Biotechnol.* **18**, 538–543 (2000).
35. Alberti, S., Gitler, A. D. & Lindquist, S. A suite of Gateway cloning vectors for high-throughput genetic analysis in *Saccharomyces cerevisiae*. *Yeast* **24**, 913–919 (2007).
36. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).
37. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
38. Szklarczyk, D. et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45** (D1), D362–D368 (2017).
39. Hassink, G. et al. TEB4 is a C4HC3 RING finger-containing ubiquitin ligase of the endoplasmic reticulum. *Biochem. J.* **388**, 647–655 (2005).
40. Loregger, A. et al. A MARCH6 and IDOL E3 ubiquitin ligase circuit uncouples cholesterol synthesis from lipoprotein uptake in hepatocytes. *Mol. Cell. Biol.* **36**, 285–294 (2015).
41. Stevenson, J., Luu, W., Kristiana, I. & Brown, A. J. Squalene mono-oxygenase, a key enzyme in cholesterol synthesis, is stabilized by unsaturated fatty acids. *Biochem. J.* **461**, 435–442 (2014).
42. Zelcer, N. et al. The E3 ubiquitin ligase MARCH6 degrades squalene monooxygenase and affects 3-hydroxy-3-methyl-glutaryl coenzyme A reductase and the cholesterol synthesis pathway. *Mol. Cell. Biol.* **34**, 1262–1270 (2014).
43. Nomura, J. et al. Neuroprotection by endoplasmic reticulum stress-induced HRD1 and chaperones: possible therapeutic targets for Alzheimer's and Parkinson's disease. *Med. Sci.* **4**, E14 (2016).
44. Joshi, V., Upadhyay, A., Kumar, A. & Mishra, A. Gp78 E3 ubiquitin ligase: essential functions and contributions in proteostasis. *Front. Cell. Neurosci.* **11**, 259 (2017).
45. Zenker, M. et al. Deficiency of UBR1, a ubiquitin ligase of the N-end rule pathway, causes pancreatic dysfunction, malformations and mental retardation (Johanson-Blizzard syndrome). *Nat. Genet.* **37**, 1345–1350 (2005); corrigendum 38, 265 (2006).
46. George, A. J., Hoffiz, Y. C., Charles, A. J., Zhu, Y. & Mabb, A. M. A comprehensive atlas of E3 ubiquitin ligase mutations in neurological disorders. *Front. Genet.* **9**, 29 (2018).
47. Mezghrani, A. et al. A destructive interaction mechanism accounts for dominant-negative effects of misfolded mutants of voltage-gated calcium channels. *J. Neurosci.* **28**, 4501–4511 (2008).
48. Manganas, L. N. et al. Episodic ataxia type-1 mutations in the Kv1.1 potassium channel display distinct folding and intracellular trafficking properties. *J. Biol. Chem.* **276**, 49427–49434 (2001).
49. Mittal, S., Dubey, D., Yamakawa, K. & Ganesh, S. Lafora disease proteins malin and laforin are recruited to aggresomes in response to proteasomal impairment. *Hum. Mol. Genet.* **16**, 753–762 (2007).
50. Atkin, T. A., Brandon, N. J. & Kittler, J. T. Disrupted in schizophrenia 1 forms pathological aggresomes that disrupt its function in intracellular transport. *Hum. Mol. Genet.* **21**, 2017–2028 (2012).
51. Crider, A., Ahmed, A. O. & Pillai, A. Altered expression of endoplasmic reticulum stress-related genes in the middle frontal cortex of subjects with autism spectrum disorder. *Mol. Neuropsychiatry* **3**, 85–91 (2017).
52. De Jaco, A., Comoletti, D., King, C. C. & Taylor, P. Trafficking of cholinesterases and neuroglins mutant proteins. An association with autism. *Chem. Biol. Interact.* **175**, 349–351 (2008).

53. De Jaco, A. et al. A mutation linked with autism reveals a common mechanism of endoplasmic reticulum retention for the alpha,beta-hydrolase fold protein family. *J. Biol. Chem.* **281**, 9667–9676 (2006).
54. De Jaco, A. et al. Neuroligin trafficking deficiencies arising from mutations in the alpha/beta-hydrolase fold protein family. *J. Biol. Chem.* **285**, 28674–28682 (2010).
55. Fujita, E. et al. Autism spectrum disorder is related to endoplasmic reticulum stress induced by mutations in the synaptic cell adhesion molecule, CADM1. *Cell Death Dis.* **1**, e47 (2010).
56. Ulbrich, L. et al. Autism-associated R451C mutation in neuroligin3 leads to activation of the unfolded protein response in a PC12 Tet-On inducible system. *Biochem. J.* **473**, 423–434 (2016).
57. El Ayadi, A., Stieren, E. S., Barral, J. M. & Boehning, D. Ubiquilin-1 and protein quality control in Alzheimer disease. *Prion* **7**, 164–169 (2013).
58. Marín, I. The ubiquilin gene family: evolutionary patterns and functional insights. *BMC Evol. Biol.* **14**, 63 (2014).
59. Safren, N. et al. Ubiquilin-1 overexpression increases the lifespan and delays accumulation of Huntingtin aggregates in the R6/2 mouse model of Huntington's disease. *PLoS One* **9**, e87513 (2014).
60. Natunen, T. et al. Relationship between ubiquilin-1 and BACE1 in human Alzheimer's disease and APD9 transgenic mouse brain and cell-based models. *Neurobiol. Dis.* **85**, 187–205 (2016).
61. Deng, H. X. et al. Mutations in UBQLN2 cause dominant X-linked juvenile and adult-onset ALS and ALS/dementia. *Nature* **477**, 211–215 (2011).
62. Osaka, M., Ito, D. & Suzuki, N. Disturbance of proteasomal and autophagic protein degradation pathways by amyotrophic lateral sclerosis-linked mutations in ubiquilin 2. *Biochem. Biophys. Res. Commun.* **472**, 324–331 (2016).
63. Teyssou, E. et al. Novel UBQLN2 mutations linked to amyotrophic lateral sclerosis and atypical hereditary spastic paraplegia phenotype through defective HSP70-mediated proteolysis. *Neurobiol. Aging* **58**, 239e11–239e220 (2017).
64. Zeng, L. et al. Differential recruitment of UBQLN2 to nuclear inclusions in the polyglutamine diseases HD and SCA3. *Neurobiol. Dis.* **82**, 281–288 (2015).
65. Hjerpe, R. et al. UBQLN2 mediates autophagy-independent protein aggregate clearance by the proteasome. *Cell* **166**, 935–949 (2016).

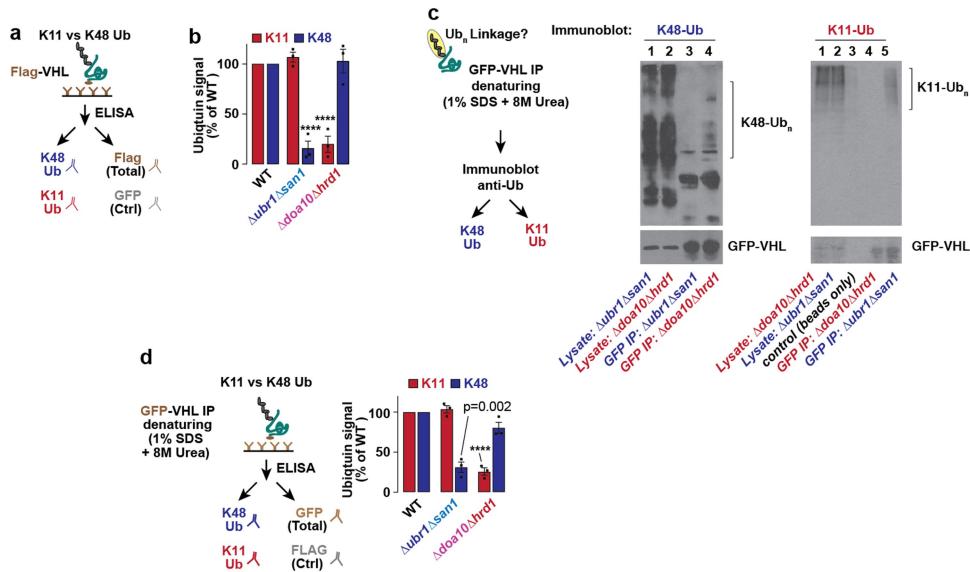


Extended Data Fig. 1 | E3 ligases work in optimal combinations to clear misfolded proteins. **a**, Assay for puncta formation distinguishes between misfolded versus natively folded proteins. WT cells expressing natively folded Ubc9^{WT}-GFP or temperature-sensitive Ubc9^{ts}-GFP from a galactose-inducible promoter for 4–6 h at 30 °C were shifted to glucose-containing medium for 1 h at 30 °C or 37 °C to shut off expression. Cells were fixed and imaged by fluorescence microscopy. 300 cells were counted per condition, and the percentage of cells with GFP-positive puncta is shown (means \pm s.e.m. from three biologically independent experiments). Only cells expressing Ubc9^{ts}-GFP showed a statistically significant change in the percentage of puncta-positive cells compared with WT (two-tailed Student's *t*-test, $****P < 0.0001$; ns, not significant). **b**, Deletion of individual E3 ligases does not increase puncta formation. Experiment performed as in panel a, but using strains with endogenous deletions of the genes shown on the x-axis. E3 ligases that have previously been implicated in PQC (as shown in Fig. 1d) are grouped to the right (QC, quality control). Bars represent mean \pm s.e.m. from three biologically independent experiments, with the exception of rad5, hel1, etp1, irc20, hel2, apc11, hrt1, tfb3, cdc24, prp19, upf1, upf3, itt1 and rad18, where bars represent the mean from two biologically independent experiments, as well as WT, where bars represent the mean \pm s.e.m. from seven biologically independent experiments. No strains showed statistically significant differences compared with WT by one-way ANOVA followed by Dunnett's multiple comparisons test. **c**, Deleting certain pairs of E3 ligases increases the stability of misfolded proteins. CHX chase was followed by immunoblot to assess the stability of GFP-VHL, CPY^{ts}-GFP, Ubc9^{ts}-TAP or Ubc9^{WT}-TAP in E3 double-deletion strains. For the WT + Bz condition, 50 μ M Bz was added to the glucose-containing medium 10 min before CHX treatment. Graphs represent densitometric quantification of bands relative to $t = 0$ (mean \pm s.e.m. from three biologically independent experiments). **d**, Multiple misfolded proteins are sequestered in the same subcellular location. *Ubr1Δsan1* or *Δdoa10Δhrd1* strains co-expressing VHL with temperature-sensitive Ubc9^{ts} (top) or natively folded Ubc9^{WT} (bottom) from galactose-inducible promoters for 5–6 h at 30 °C were shifted to glucose-containing medium for 1 h at 37 °C. Fluorescence microscopy images are representative of at least 100 cells from each of three biologically independent experiments. **e**, Deletion of certain pairs of E3s increases puncta formation. Experiment performed as in panel a, but in strains with endogenous deletions of pairs of E3 genes. Each right-hand panel shows experiments in which cells were shifted to 37 °C for the 1 h of galactose shut-off. Bars represent means \pm s.e.m. from three biologically independent experiments. Strains for which statistically significant differences were observed by one-way ANOVA followed by Dunnnett's multiple comparisons test compared with WT are indicated with the adjusted *P* value or with $****$ for $P < 0.0001$. **f**, Overexpressing a single E3 ligase does not compensate for the loss of others. Ubr1, San1 or Hrd1 were overexpressed alongside GFP-VHL in the indicated strains. The rest of the experiment was performed as in a. Bars represent means <



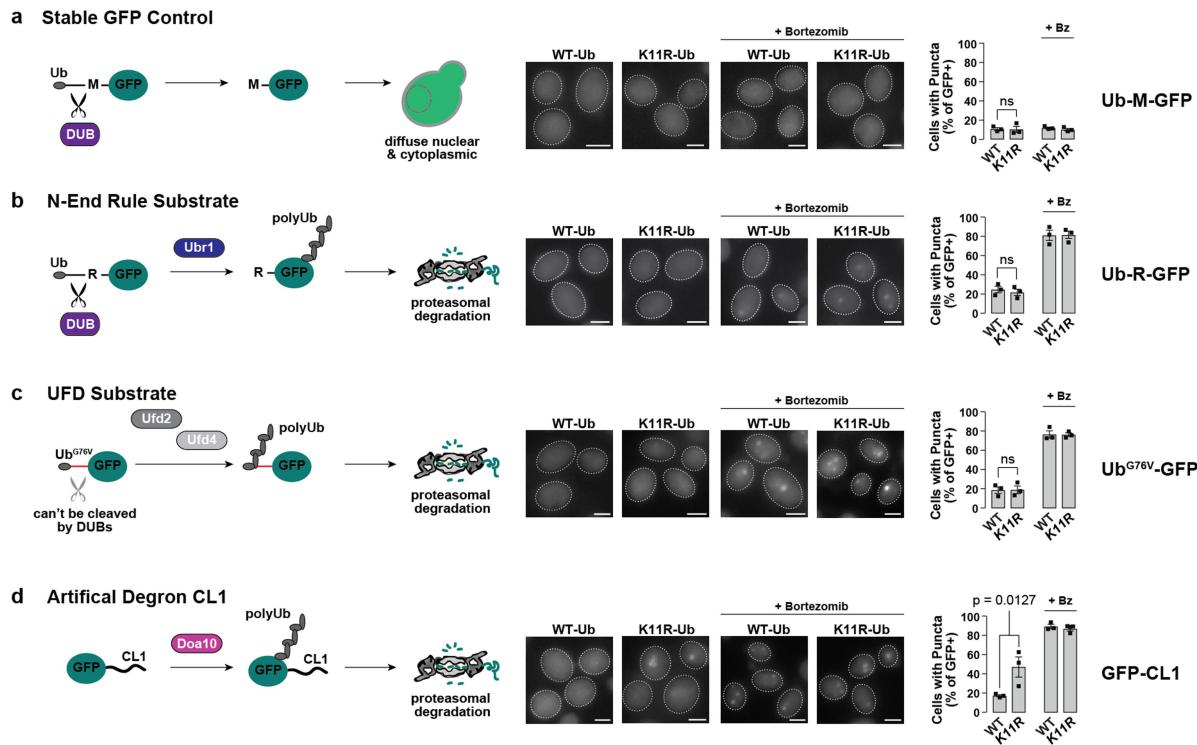
Extended Data Fig. 2 | San1 forms a complex with Doa10 but not with Hrd1. **a, b**, San1–V5His6 co-immunoprecipitates with Doa10–GFP but not with Hrd1–GFP. Yeast cells co-expressing Doa10–GFP (a) or Hrd1–GFP (b) from their endogenous promoters with San1–V5His6 from a galactose-inducible promoter for 16 h were shifted to 37 °C for 1 h, and immediately lysed by cryo-grinding. Native complexes were immunoprecipitated with GFP-Trap-MA nanobodies before

immunoblotting with the indicated antibodies. Immunoblots are representative of three biologically independent experiments. **c, d**, Flag–Ubr1 does not co-immunoprecipitate with Doa10–GFP or Hrd1–GFP. The experiment was performed as in panel a and b, but with cells expressing Flag–Ubr1 (from the constitutive ADH promoter) instead of San1–V5His6. Immunoblots are representative of three biologically independent experiments.



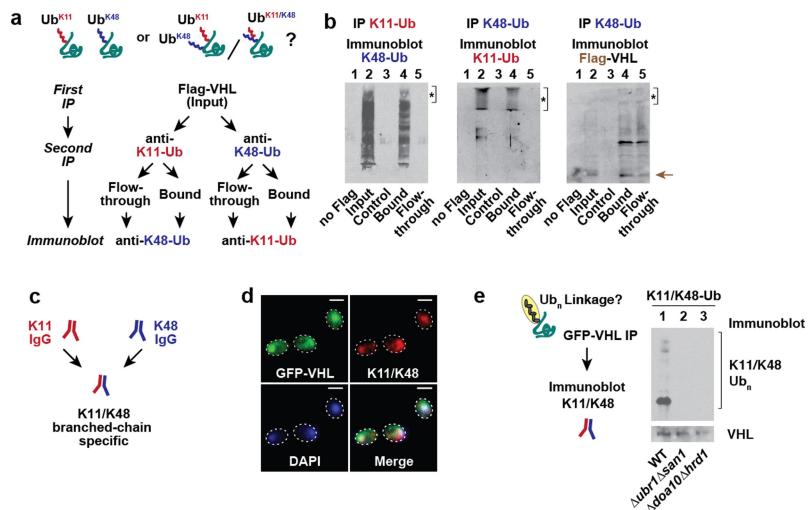
Extended Data Fig. 3 | K48–Ub and K11–Ub linkages are reduced in $\Delta ubr1\Delta san1$ and $\Delta doa10\Delta hrd1$ strains, respectively. **a**, Diagram showing the Ub-linkage ELISA used to quantify Ub linkages. Flag–VHL from a yeast lysate was immunoprecipitated in an anti-Flag-conjugated 96-well plate (using four wells per sample), and incubated with antibodies against GFP (negative control), Flag, K11–Ub, or K48–Ub. Following incubation with a secondary antibody (anti-rabbit-HRP), the strength of each signal was detected by electrochemiluminescence at 450 nm. To quantify the K11–Ub or K48–Ub linkages on Flag–VHL, we subtracted the anti-K11 or anti-K48 signal from the negative control (anti-GFP) and normalized to the total Flag–VHL signal for each sample. **b**, Ub-linkage ELISA confirms that K48–Ub and K11–Ub linkages are reduced on Flag–VHL in $\Delta ubr1\Delta san1$ and $\Delta doa10\Delta hrd1$ strains, respectively. WT or E3 double-deletion strains expressing Flag–VHL at 30 °C for 4–6 h were lysed after 1 h Bz treatment, also at 30 °C. Ub-linkage ELISA was then performed as described in **a**. Bars represent Flag-normalized values from each strain (mean \pm s.e.m. from three biologically independent experiments), expressed as a proportion of the Flag-normalized WT

values. Strains with statistically significant differences compared with WT by one-way ANOVA followed by Dunnett's multiple comparisons test are indicated (****P < 0.001). **c**, GFP–VHL denaturing immunoprecipitation (1% SDS + 8 M urea) followed by immunoblot for K48–Ub or K11–Ub in WT or E3 double-deletion strains. Immunoblots are representative of three independent experiments. **d**, Relative amounts of K11–Ub and K48–Ub linkages present on GFP–VHL in $\Delta ubr1\Delta san1$ or $\Delta doa10\Delta hrd1$ strains compared with WT. WT or E3 double-deletion strains expressing GFP–VHL at 30 °C for 5–6 h were lysed in denaturing conditions (1% SDS + 8 M urea) after 1 h Bz treatment, also at 30 °C. Ub-linkage ELISA was then performed using GFP-multiTrap plates. Bars represent GFP-normalized values from each strain (means \pm s.e.m. from three biologically independent experiments) expressed as a proportion of the GFP-normalized WT values. Strains for which statistically significant differences were observed by one-way ANOVA followed by Dunnett's multiple comparisons test compared with WT are indicated with the adjusted P value, or with **** for P < 0.0001.



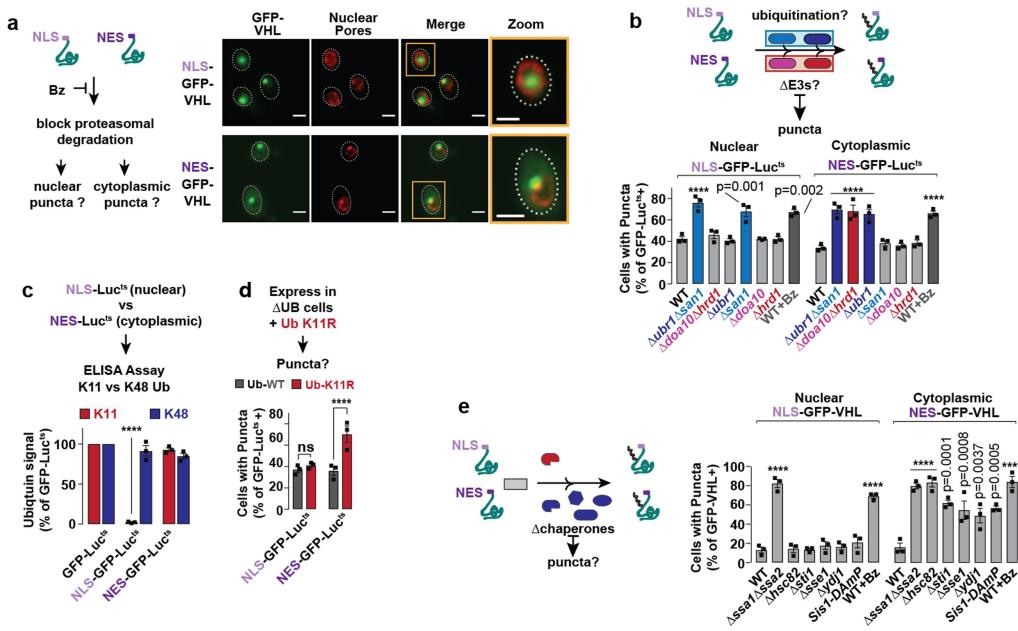
Extended Data Fig. 4 | K11-Ub linkages are not necessary for proteasomal degradation of all cytoplasmic substrates. **a–d**, WT or Ub^{K11R} cells expressing stable Ub-M-GFP (a), the N-end-rule substrate Ub-R-GFP (b), the ubiquitin fusion degradation (UFD) substrate Ub^{G76V}-GFP (c) or GFP fused to the artificial degron CL1 (d) from galactose-inducible promoters for 4–6 h at 30 °C were shifted to glucose-containing medium for 1 h at 30 °C or 37 °C to shut off expression. Cells were fixed and imaged by fluorescence microscopy. 300 cells were counted per

condition, and the percentage of cells with GFP-positive puncta is shown (mean ± s.e.m. from three biologically independent experiments). There was a statistically significant increase in puncta compared with WT when GFP-CL1 (which contains a short amphipathic CL1 helix that could mimic a partially unfolded protein) was expressed in Ub^{K11R} cells, as judged by two-tailed Student's *t*-test ($P=0.0127$). The differences for all other substrates were not significant (ns, $P>0.05$). DUB, deubiquitinating enzyme, which cleaves Ub from Ub-M-GFP or Ub-R-GFP.



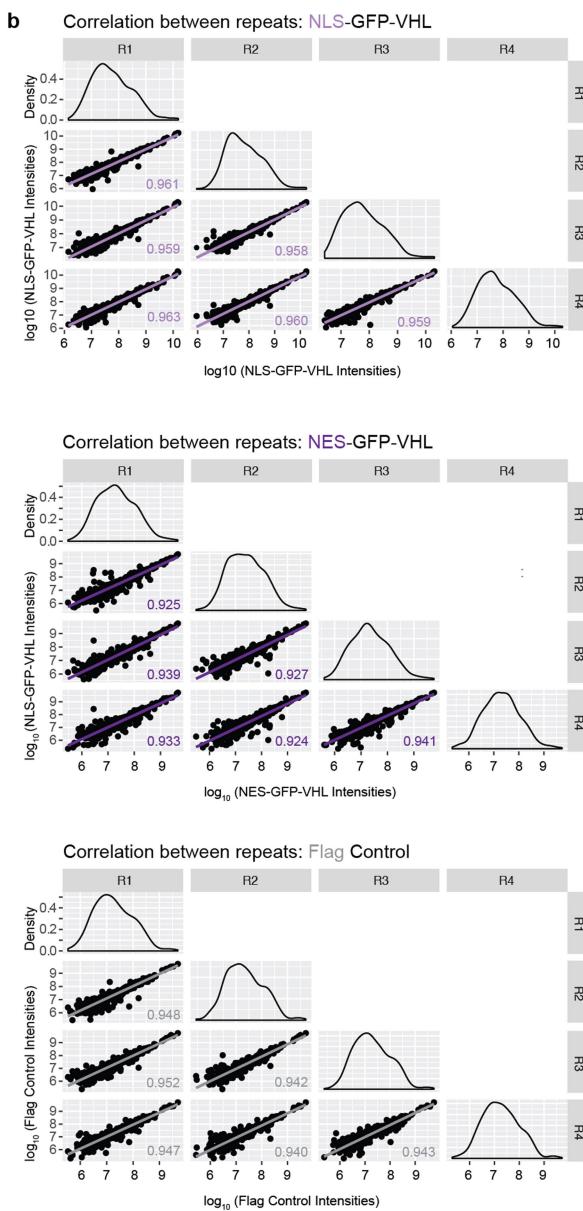
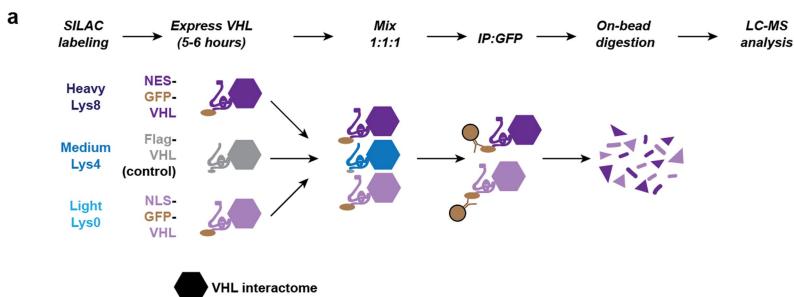
Extended Data Fig. 5 | Misfolded VHL is modified with branched K11/K48 ubiquitin chains. **a, b**, Both K11-Ub and K48-Ub linkages are present on the same VHL molecule. **a**, This experiment was designed to determine whether both K48-Ub and K11-Ub linkages are present in the same VHL population. Sequential immunoprecipitation was carried out, first with anti-Flag antibody, then with an anti-K11-Ub or anti-K48-Ub antibody. The resulting negative control ('no Flag', with mock Flag plus K11 or K48 immunoprecipitation with lysate from cells expressing GFP-VHL instead of Flag-VHL), bead control ('Control', with no K11-Ub or K48-Ub antibody), 'Bound' and 'Flow-through', in addition to samples with just the first Flag immunoprecipitation (Input), were subjected to SDS-PAGE and immunoblotted for the presence of the other Ub linkage (**b**). Immunoblots representative of three biologically independent experiments are shown. The arrow indicates the size of un-ubiquitinated Flag-VHL. The asterisks indicate proteins in the stacking gel that did not enter the resolving gel. **c**, This bispecific

anti-K11/K48-Ub antibody was designed to bind ubiquitin chains with K11 and K48 linkages branching off the same ubiquitin moiety. **d**, Misfolded VHL co-localizes with K11/K48-Ub chains. WT cells expressing GFP-VHL from a galactose-inducible promoter for 4–6 h at 30 °C were shifted to glucose-containing medium with 50 μM bortezomib for 1 h to shut off expression. Cells were fixed, spheroplasted and detergent permeabilized before immunostaining with an antibody designed to recognize ubiquitin that had K11 and K48 linkages emanating from the same moiety (K11/K48). Confocal fluorescence microscopy images are representative of at least 100 cells from each of three biologically independent experiments. Scale bars represent 2 μm. **e**, VHL is modified with branched K11/K48-Ub chains. GFP-VHL denaturing immunoprecipitation was followed by immunoblot for K11/K48-Ub or GFP (VHL) in WT or E3 double-deletion strains. Immunoblots representative of three biologically independent experiments are shown.



Extended Data Fig. 6 | Nuclear and cytoplasmic proteins require different PQC pathways for clearance. **a**, NLS-GFP-VHL and NES-GFP-VHL form a single punctum in the nucleus or cytoplasm, respectively, upon proteasome inhibition. WT cells expressing NLS-GFP-VHL or NES-GFP-VHL from a galactose-inducible promoter for 4–6 h at 30 °C were shifted to glucose-containing medium with 50 µM Bz for 1 h at 30 °C to shut off GFP-VHL expression. Fixed and spheroplastral cells were immunostained for the nuclear pore complex protein Nsp1 (red) before imaging by fluorescence microscopy. Representative cells from three biologically independent repeats are shown. **b–d**, Misfolded luciferase^{ts} (Luc^{ts}) confined to the nucleus can be cleared by San1-mediated K48-linked ubiquitination. **b**, The increase in the percentage of cells containing puncta of NLS-GFP-Luc^{ts} and NES-GFP-Luc^{ts} across the E3 single- and double-deletion strains is similar to the pattern observed with NLS-GFP-VHL and NES-GFP-VHL in Fig. 4. Shown is the percentage of cells (mean ± s.e.m. from three biologically independent experiments, each with $n = 300$) containing NLS-GFP-Luc^{ts} or NES-GFP-Luc^{ts} puncta in WT, single- or double-deletion strains after 4–6 h expression of the protein at 30 °C followed by 1 h shut-off at 37 °C. Strains for which statistically significant differences were observed by one-way ANOVA followed by Dunnett's multiple comparisons test compared with WT are indicated with the adjusted P value, or with *** for $P < 0.0001$. **c**, Misfolded nuclear Luc^{ts} has severely reduced K11-Ub linkages

(**** $P < 0.0001$ by one-way ANOVA followed by Dunnett's multiple comparisons test). Ubiquitin-linkage ELISA was performed on lysates of WT yeast expressing NLS-, NES- or unaltered GFP-Luc^{ts} at 37 °C as described in Extended Data Fig. 2c, but in GFP-multiTrap 96-well plates instead of anti-Flag-conjugated 96-well plates. Anti-Flag was used instead of anti-GFP as the ELISA negative control. Bars represent means ± s.e.m. from three biologically independent experiments. **d**, Misfolded luciferase^{ts} confined to the nucleus does not require K11R-Ub linkages for clearance. The experiment was performed as in panel **b**, but with yeast strains expressing WT or mutant K11R-Ub as their sole source of ubiquitin. 300 cells were counted per condition, and the percentages of cells with GFP-positive puncta are shown in (means ± s.e.m. from three biologically independent experiments). Only NES-GFP-Luc^{ts} had a statistically significant change in puncta-positive cells in the K11R strain when compared with WT (one-way ANOVA followed by Dunnett's multiple comparisons test; *** $P < 0.0001$; ns = $P > 0.05$). **e**, VHL confined to the nucleus (NLS) or cytoplasm (NES) requires different chaperones for clearance. The experiment was performed as in panel **b**, but with the indicated chaperone-deletion strains. Bars represent means ± s.e.m. from three biologically independent experiments. Strains for which statistically significant differences were observed compared with WT by one-way ANOVA followed by Dunnett's multiple comparisons test are indicated with the adjusted P value, or with *** for $P < 0.0001$.



c Enriched PQC proteins [\log_2 (VHL/Control) > 0.5]

	NLS-GFP-VHL	NES-GFP-VHL
Proteasome		
20S core particle	Pre10	Pre5 Pre6 Pre9
19S regulatory particle	Rpt6 Rpt1 Rpn2	Rpt6
		Rpn3 Rpn6
Chaperones		
<i>Hsp70</i> family	Ssa1 Ssa2 Ssa4 Ssb2	Ssa1 Ssa2 Ssa4 Ssb2
<i>Hsp90</i> family		Hsp82
<i>Hsp40</i> family		Sis1
<i>TRiC/CCT</i>	Cct5 Cct3 Cct4 Cct6 Cct7 Cct8	Cct5
Thioredoxins	Trx1 Trx2 Tsa1	Trx1 Trx2 Tsa1
14-3-3		Bmh1
Degradation machinery		Uba1 Cdc48 Crm1

d Selected enriched PQC pathways (adjusted $p < 0.05$)

	NLS-GFP-VHL	NES-GFP-VHL
GO-Biological Process		
Proteasome assembly	0.0189	0.0286
Protein folding	n.s.	0.0121
GO-Molecular Function		
Unfolded protein binding	0.0076	0.0047
GO-Cellular Component		
Chaperonin-containing T-complex	2.87E-09	n.s.
KEGG Pathways		
Proteasome	0.00401	6.27E-06
PFAM Protein Domains		
<i>TCP-1/cpn60 chaperonin</i> family	2.49E-06	n.s.
<i>Hsp70</i> protein	0.00229	0.00787

Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Mass spectrometry of the VHL interactome identifies distinct PQC circuitries for nuclear and cytoplasmic VHL.
a. Triple SILAC-base mass spectrometry of VHL immunoprecipitates. WT yeast cells transfected with one of NLS–GFP–VHL, NES–GFP–VHL or Flag–VHL were grown overnight at 30 °C in raffinose-synthetic media supplemented with light Lys0, heavy Lys8 or medium Lys4, respectively. Growth of VHL was induced in galactose for 4–5 h before shut off in glucose for 90 min. Next, 1.5 mg of protein from each of the three lysed samples were mixed before immunoprecipitation using GFP-TRAP_MA magnetic bead on-bead restriction digestion and peptide clean-up. Peptides were identified using liquid-chromatography/mass-spectrometry analysis before analysis using MaxQuant. **b.** Strong correlation between the four biological repeats (R1–R4). Raw intensities for light (NLS–GFP–VHL; top), heavy (NES–GFP–VHL; middle) and medium (VHL–Flag control, bottom) were \log_{10} -transformed and plotted as scatterplot matrices. The Pearson correlation coefficient for each pairwise comparison is indicated, and the density distribution of intensities within each repeat is shown in the diagonal axis of the matrices. **c.** Enriched

PQC proteins in NLS–GFP–VHL and NES–GFP–VHL interactomes. Normalized median light/medium (NLS–GFP–VHL) and heavy/medium (NES–GFP–VHL) SILAC ratios were \log_2 -transformed. Proteins with \log_2 (SILAC ratio) of greater than 0.5 were considered as enriched, yielding 49 and 56 proteins for the NLS and NES interactomes, respectively. Enriched proteins known to play a role in PQC are shown. Both nuclear and cytoplasmic VHL share enrichments in proteasomal subunits, the Hsp70 chaperones Ssa1, Ssa2, Ssa4 and Ssb2, and the thioredoxins Trx1, Trx2 and Tsa1 (previously implicated in misfolded-protein management). All enriched proteins are shown in Extended Data Table 1. **d.** Enriched PQC pathways in NLS–GFP–VHL and NES–GFP–VHL interactomes. The enriched proteins from each interactome (median values from four biologically independent experiments) were subjected to pathway analysis to search for enriched GO terms, KEGG pathways and PFAM protein domains in either interactome using the STRING database. Selected enriched PQC pathways are shown ($P < 0.05$ using Fisher's exact test followed by Benjamini–Hochberg multiple testing correction).

Extended Data Table 1 | Protein and pathways enriched in nuclear and cytoplasmic interactomes

Enriched proteins [$\log_2(\text{NLS-VHL}/\text{Control}) > 0.5$]

Fasta headers	$\log_2(\text{Median NLS}/\text{Control})$
URA3	2.077
CCT5	1.550
SSA2	1.334
SSA4	1.239
MKT1	1.208
TSA1	1.203
TRX2	1.035
CCT8	0.952
ARO1	0.923
MDN1	0.902
RPT6	0.879
CCT7	0.873
TUB1	0.867
CCT3	0.841
GCD6	0.837
TRX1	0.808
BGL2	0.790
VMA2	0.775
AAH1	0.773
NEW1	0.759
RPA135	0.745
SSA1	0.745
YNL134C	0.728
URA7	0.710
PRE10	0.699
EFT1;EFT2	0.688
URA2	0.684
RPT1	0.670
NUG1	0.657
HTS1	0.649
SSB2	0.642
SAM1	0.630
RPB2	0.615
KRE33	0.605
RPN2	0.578
FAS2	0.561
ADE6	0.552
MIS1	0.550
YEF3	0.545
CCT6	0.540
CPA2	0.524
CCT4	0.520
ADE3	0.516
RRP5	0.514
LEU1	0.510
CRM1	0.508
GCN1	0.507
HIS1	0.507
HIS4	0.504

Enriched proteins [$\log_2(\text{NES-VHL}/\text{Control}) > 0.5$]

Fasta headers	$\log_2(\text{Median NES}/\text{Control})$
YNL134C	3.229
DBP5	2.501
RK11	2.494
URA3	2.319
SSA4	2.166
SHM1	1.835
SEC14	1.727
TSA1	1.234
SSA2	1.234
IMD3;IMD2	1.213
SSA1	1.199
TRX2	1.054
GDH1	0.984
BGL2	0.964
TPI1	0.880
RPN3	0.876
CCT5	0.820
CIT1	0.803
TIF4631	0.792
CPA2	0.776
RRP5	0.771
PRE5	0.731
IPP1	0.731
MET17	0.730
DYS1	0.702
SIS1	0.695
TRX1	0.675
MMF1	0.672
ENO2	0.672
PGI1	0.670
RPT6	0.665
PRE9	0.658
HOM6	0.645
HXK2	0.628
ERG10	0.625
BMH1	0.625
UBA1	0.616
PGK1	0.612
HTS1	0.608
YPT52	0.607
SSB2	0.607
URA2	0.598
RPN6	0.597
FPR4	0.592
TKL1	0.578
PRE6	0.571
RNR4	0.561
CDC48	0.558
SDH1	0.535
HSP82	0.526
TUB1	0.520
YPL225W	0.515
TAL1	0.508
LEU1	0.506
GPM1	0.503

Enriched proteins in the NLS-GFP-VHL (left) and NES-GFP-VHL (right) interactomes. Normalized median light/medium (NLS-GFP-VHL) and heavy/medium (NES-GFP-VHL) SILAC ratios from four biologically independent experiments were \log_2 -transformed. Proteins with $\log_2(\text{SILAC ratio})$ greater than 0.5 were considered as enriched, yielding 49 and 56 proteins for the NLS and NES interactomes, respectively.

Extended Data Table 2 | Human homologues of the ubiquitination machinery characterized here are associated with a range of diseases

Yeast gene	Human orthologue	Associated Diseases	PQC in disease pathology?
Doa10	MARCH6/TEB4	Cri-du-chat syndrome (same chromosomal region) ³⁹ ; Lipidogenesis imbalance ⁴⁰⁻⁴²	-
Hrd1	HRD1	Alzheimer's Disease Parkinson's Disease	⁴³ (Review)
	Gp78/AMFR	Cancer Cystic fibrosis ALS Parkinson's Disease Huntington's Disease prion disorders	⁴⁴ (Review)
Ubr1	UBR1	Johanson-Blizzard Syndrome ⁴⁵	⁴⁵
	UBR4	Episodic Ataxia Type 8 ⁴⁶	Unclear (yes for related Types 1 & 2) ^{47,48}
	UBR5	Adult Myoclonal Epilepsy ⁴⁶	Unclear (yes for related Lafora Myoclonal Epilepsy) ⁴⁹
	UBR7	Autism Spectrum Disorder ⁴⁶	⁵⁰⁻⁵⁶
San1	No clear human orthologue	-	-
Dsk2	UBQLN1	Alzheimer's Disease Huntington's Disease	^{57,58} (Reviews), ^{59,60}
	UBQLN2	ALS Frontotemporal Dementia Huntington's Disease	⁶¹⁻⁶⁵

Literature-based evidence for disease pathology being directly related to PQC is indicated in the last column, where applicable. References cited are³⁹⁻⁶⁵.

The metabolite dimethylsulfoxonium propionate extends the marine organosulfur cycle

Kathleen Thume¹, Björn Gebser¹, Liang Chen², Nils Meyer¹, David J. Kieber^{2*} & Georg Pohnert^{1,3*}

Algae produce massive amounts of dimethylsulfoniopropionate (DMSP), which fuel the organosulfur cycle^{1,2}. On a global scale, several petagrams of this sulfur species are produced annually, thereby driving fundamental processes and the marine food web¹. An important DMSP transformation product is dimethylsulfide, which can be either emitted to the atmosphere^{3,4} or oxidized to dimethylsulfoxide (DMSO) and other products⁵. Here we report the discovery of a structurally unusual metabolite, dimethylsulfoxonium propionate (DMSOP), that is synthesized by several DMSP-producing microalgae and marine bacteria. As with DMSP, DMSOP is a low-molecular-weight zwitterionic metabolite that carries both a positively and a negatively charged functional group. Isotope labelling studies demonstrate that DMSOP is produced from DMSP, and is readily metabolized to DMSO by marine bacteria. DMSOP was found in near nanomolar amounts in field samples and in algal culture media, and thus represents—to our knowledge—a previously undescribed biogenic source for DMSO in the marine environment. The estimated annual oceanic production of oxidized sulfur from this pathway is in the teragram range, similar to the calculated dimethylsulfide flux to the atmosphere³. This sulfoxonium metabolite is therefore a key metabolite of a previously undescribed pathway in the marine sulfur cycle. These findings highlight the importance of DMSOP in the marine organosulfur cycle.

The marine organosulfur cycle is fuelled by small sulfur-containing zwitterionic osmolytes that are primarily produced by planktonic algae. The main metabolite of this class, DMSP, is produced in the impressive amounts of 2 petagrams (2×10^8 tons) sulfur annually¹. Cellular DMSP serves important physiological functions in marine algae that include, but are not limited to, acting as an osmolyte, a cryoprotectant and an antioxidant^{6,7}. Enzymatic lysis of DMSP by DMSP lyases in bacteria and algae yields acrylate and dimethylsulfide (DMS)⁸. Volatile DMS is the main source of organosulfur in the atmosphere; and with an annual flux of approximately 30 teragrams of sulfur³, DMS has been proposed to affect cloud formation and regulate climate⁴. Dissolved DMSP arising from exudation, grazing, viral lysis and cell mortality serves as substrate for marine microorganisms^{7,9,10}. In surface waters, substantial quantities of dissolved DMSP and DMS can be detected, but often the concentration of dissolved DMSO exceeds the concentration of each of these two species^{5,11}. DMSO is mainly produced from bacterial and photochemical DMS oxidation¹², but algal sources of DMSO may also be important¹³. Common pelagic bacteria use monooxygenases to oxidize DMS to DMSO¹⁴, a process that may serve as an energy source¹⁵. Here we report on the identification of the zwitterionic metabolite DMSOP, which is widely distributed in phytoplankton and also produced by marine bacteria. This metabolite is the substrate of a previously undescribed marine pathway for DMSO production (Fig. 1).

Zwitterionic metabolites, such as DMSP, are difficult to quantify directly and much information on their chemistry and ecology is based on indirect methods. We recently developed an analytical method to directly detect zwitterionic metabolites^{16,17}, and observed discrepancies

between our analytical data and previous determinations of DMSP and DMSO in plankton samples. We undertook an in-depth survey to determine whether additional metabolites could explain this observation, and consistently detected a compound with similar polarity to DMSP in all main classes of microalgae (Fig. 2 and Table 1). The electrospray ionization high-resolution mass spectrum in positive ionization mode of this metabolite at $m/z = 151.0426$ was consistent with the formula $C_5H_{11}O_3S$ (calculated $m/z = 151.0423$), and the isotope peak at $m/z = 153.0378$ (calculated $m/z = 153.0380$) confirmed the presence of a sulfur atom in the structure. A fragment ion $m/z = 79.0210$ was detected by tandem mass spectrometry (MS/MS) that was attributed to protonated DMSO and a fragment at $m/z = 73.0283$ corresponded to protonated acrylic acid (Fig. 2 and Extended Data Fig. 1). On the basis of the mass spectral data, the signal was tentatively assigned as

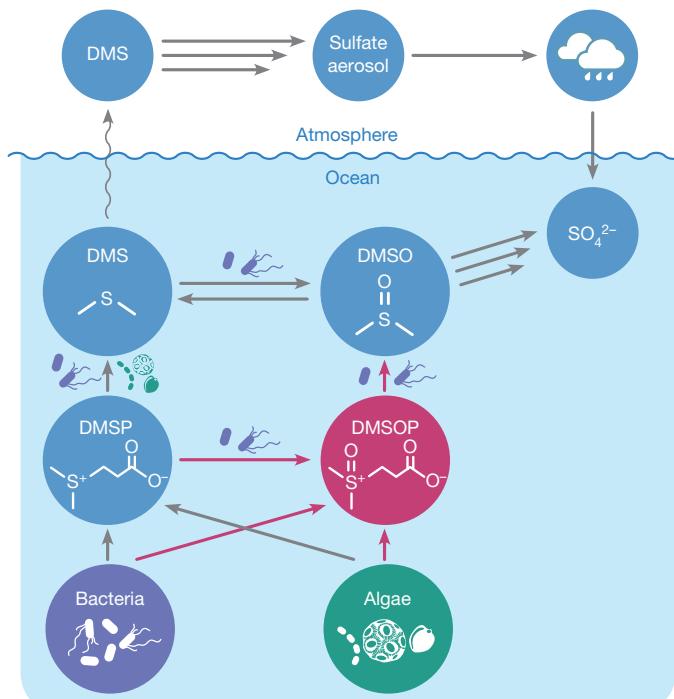


Fig. 1 | Simplified, revised marine sulfur cycle. DMSOP and the transformations labelled with red arrows extend the established marine sulfur cycle. DMSOP is produced in eukaryotic microalgae (green) as well as in bacteria (purple). Bacteria metabolize DMSOP and therefore contribute to the marine DMSO pool. The established DMSP-based part of the sulfur cycle is indicated with grey arrows. DMSP is formed by marine algae and bacteria. It is then cleaved by algal and bacterial DMSP lyases to DMS and acrylate (not shown). The subsequent biological and photochemical oxidation of DMS to DMSO, sulfate and other products can occur within algae, bacteria, in the seawater and the atmosphere.

¹Institute for Inorganic and Analytical Chemistry, Bioorganic Analytics, Friedrich Schiller University Jena, Jena, Germany. ²Department of Chemistry, State University of New York, College of Environmental Science and Forestry, Syracuse, NY, USA. ³Max Planck Institute for Chemical Ecology, Jena, Germany. *e-mail: djkieber@esf.edu; Georg.Pohnert@uni-jena.de

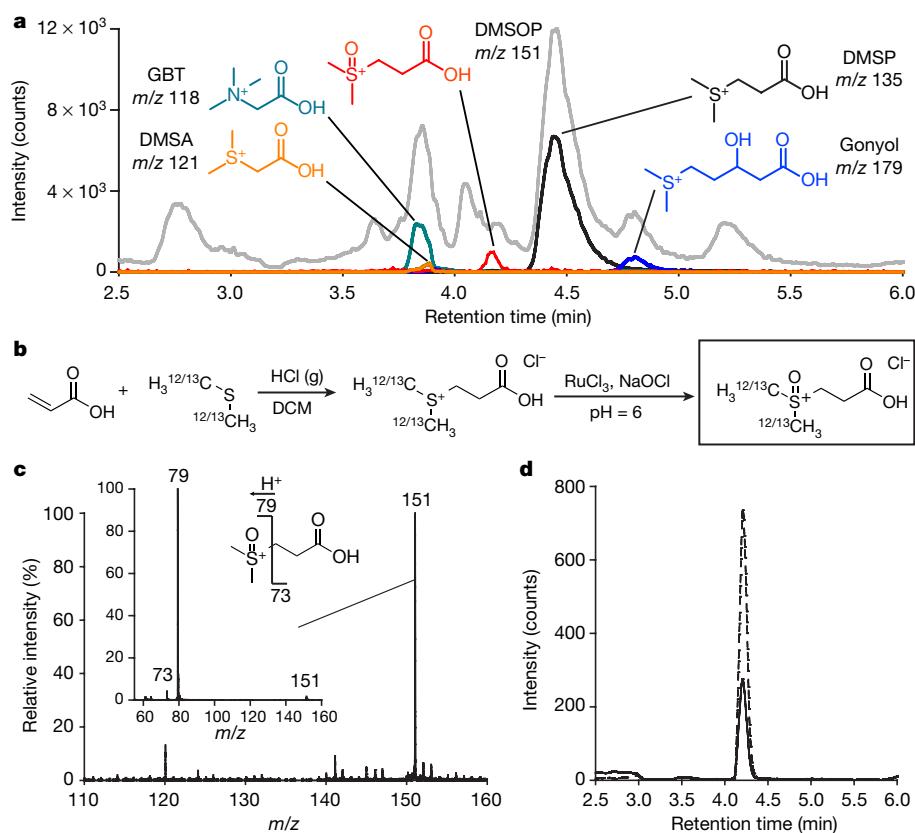


Fig. 2 | Detection and structural elucidation of DMSOP.

a, Chromatographic profile of zwitterionic metabolites from a *P. minimum* culture, separated using ultra-high-pressure liquid chromatography (UHPLC) with detection by electrospray ionization mass spectrometry (ESI-MS). The total ion current is shown in grey. The metabolites glycine betaine (GBT, cyan), dimethylsulfoniacetate (DMSA, orange), DMSP (black) and gonyol (blue) were assigned according to a previous study¹⁶.

the sulfoxonium species DMSOP. To obtain a reference compound, DMSOP was synthesized by RuCl_3 /sodium hypochlorite-mediated oxidation of DMSP, and the structure was confirmed by NMR and MS/MS (Fig. 2 and Extended Data Figs. 1, 2). When this authentic standard was added to an algal extract and submitted to liquid chromatography–MS, it coeluted with the unknown sulfur-containing metabolite, therefore unambiguously proving the identity of this highly unusual compound as DMSOP (Fig. 2d). To our knowledge, only one natural product containing the dimethylsulfoxonium moiety—(2-hydroxyethyl) dimethylsulfoxonium chloride, the causative agent for Dogger Bank itch from the marine bryozoan *Alcyonium gelatinosum*¹⁸ and the marine sponge *Theonella* aff. *mirabilis*¹⁹—has

The ion trace of DMSOP, red, is shown at a tenfold magnification.

b, Synthesis of authentic (labelled) DMSOP. **c**, Mass spectrum and tandem mass spectrum (inset) of DMSOP with characteristic fragments. **d**, UHPLC profile monitoring $m/z = 151$ of an extract of *P. parvum* (solid line) and the same extract treated with synthetic DMSOP in roughly equal amounts (dashed line), the experiment was repeated three times with varying concentrations of synthetic DMSOP to confirm co-elution.

been reported to date. Therefore, the highly polar zwitterionic DMSOP represents a metabolite of a nearly unexplored structural family.

The bloom-forming dinoflagellate *Prorocentrum minimum*, the haptophytes *Prymnesium parvum*, *Isochrysis galbana* and *Emiliania huxleyi*, the diatom *Skeletonema costatum*, and other screened diatoms and dinoflagellates all produce DMSOP (Table 1 and Extended Data Table 1) at micromolar to millimolar cellular concentrations, corresponding to 0.13–1.2% of DMSP in the algae (Table 1). DMSOP production in axenic cultures of *I. galbana* and *P. parvum* (Table 1 and Extended Data Fig. 3) confirms that phytoplankton are an oceanic source of DMSOP. The metabolite is also released into the culture medium, with concentrations of up to 0.8 ± 0.2 nM detected in a stationary axenic *P. parvum* culture.

Because marine heterotrophic bacteria biosynthesize DMSP²⁰, we investigated the possibility that DMSOP might also be a bacterial metabolite. Indeed, the DMSP-producer *Pelagibaca bermudensis* contained DMSOP (0.32 ± 0.049 pmol μg^{-1} protein, $n = 3$, approximately 0.1% of DMSP). Therefore, similar to DMSP, the oxidized sulfoxonium zwitterion has both a eukaryotic and bacterial origin. This underscores its likely universal distribution in oceanic surface waters. Consistent with this supposition, DMSOP was detected at multiple coastal sites in the northwest Pacific Ocean, northwest Atlantic Ocean, Arctic Ocean and Mediterranean Sea with an average concentration of 0.14 ± 0.18 nM. At all sampled stations, DMSOP was above the 0.01 nM limit of detection (Extended Data Fig. 6 and Extended Data Table 2). On average, DMSOP accounted for 0.22% of DMSP in field samples. This value is consistent with, but at the lower end of, the concentration range observed in culture (see above). On the basis of these findings and compared to the annual DMSP production equivalent

Table 1 | Survey of zwitterionic metabolites and results from quantification of DMSP and DMSOP

Species	GBT	DMSA	Gonyol	n_{DMSP} (fmol per cell)	n_{DMSOP} (fmol per cell)	c_{DMSOP} (mM)
<i>P. minimum</i>	+	+	+	304.5 ± 61.2	3.66 ± 1.23	0.46
<i>P. parvum</i>	+	–	+	16.2 ± 4.4	0.029 ± 0.005	0.13
<i>P. parvum</i> (axenic)	+	–	+	17.4 ± 2.6	0.023 ± 0.007	0.1
<i>S. costatum</i>	+	–	–	6.56 ± 2.06	0.029 ± 0.005	0.19
<i>E. huxleyi</i>	+	–	+	4.83 ± 0.57	0.029 ± 0.013	1.5
<i>I. galbana</i>	+	–	+	4.69 ± 0.27	0.017 ± 0.003	0.24

Presence or absence of the indicated metabolites is shown, as well as DMSP and DMSOP concentrations. Amounts per cell (n) and cellular concentration (c) are given. Data are mean \pm s.d. of $n = 3$ independent samples. The limit of detection is 0.1 μM for glycine betaine (GBT), 0.5 μM for dimethylsulfoniacetate (DMSA), 1 μM for gonyol and 0.08 μM for DMSOP. Cell volumes for determination of intracellular DMSOP concentration are given in the Methods.

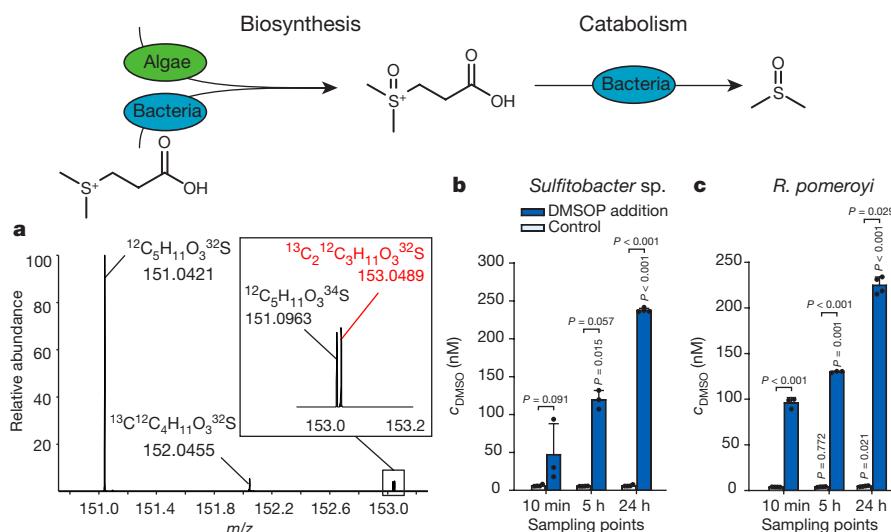


Fig. 3 | Biosynthesis and catabolism of DMSOP. **a**, High-resolution mass spectrum of DMSOP obtained from *P. bermudensis* incubated for 24 h with $^{13}\text{C}_2$ -labelled DMSP (Fig. 2). The peak labelled in red represents $^{13}\text{C}_2$ -labelled DMSOP, the natural DMSOP isotopes are shown in black (see also Extended Data Table 3). **b, c**, DMSO release (concentration (*c*) given as mean \pm s.d.) of the bacteria *Sulfitobacter* sp. and *R. pomeroyi*

to 2 petagrams sulfur per year, the corresponding estimated DMSOP sulfur flux is in the teragram range¹. This sulfur flux through DMSOP is in the same order of magnitude as the total DMS flux to the atmosphere³ (Fig. 1).

We synthesized isotopically labelled DMSOP and DMSP to study the biosynthesis and catabolism of DMSOP in *P. bermudensis* (Fig. 2b). When $^{13}\text{C}_2$ -DMSP (with labelled methyl groups at the sulfur) was added to batch cultures of *P. bermudensis*, high-resolution MS analysis revealed the formation of $^{13}\text{C}_2$ -DMSOP, with incorporation rates of $3.7 \pm 0.6\%$ after 18 h (Fig. 3 and Extended Data Table 3). Abiotic $^{13}\text{C}_2$ -DMSP oxidation to $^{13}\text{C}_2$ -DMSOP was not observed in the medium control. Similarly, no singly labelled ^{13}C -DMSOP ($m/z = 152.0457$) was detected above the intensity of the naturally occurring isotope peak, which rules out an initial DMSP demethylation, subsequent oxidation to the sulfoxide and remethylation (Extended Data Table 3). This makes it likely that enzymatic oxidation of the positively charged sulfur in DMSP was catalyzed by a hitherto unknown enzyme. The direct oxidation of DMSP to DMSOP is also consistent with previous suggestions that DMSP is involved in antioxidant processes as a consequence either of the constitutively high cellular DMSP concentrations in marine algae²¹ or the upregulation of cellular DMSP during oxidative stress⁶. Cellular DMSOP concentrations increased nearly 300% in batch cultures of *I. galbana* during the late exponential phase/stationary phase, corresponding to increased oxidative stress indicated by a decrease in the photosynthetic efficiency F_v/F_m (Extended Data Fig. 3). DMSP cellular concentrations changed very little during the growth of *I. galbana*. Because of the constitutively high DMSP concentration, this finding is consistent with the supposition that DMSP is a de facto antioxidant²¹, resulting in increased oxidative production of DMSOP from DMSP with increasing oxidative stress.

DMSOP is stable in 0.2- μm -filtered seawater at room temperature over several weeks (Extended Data Fig. 4). However, microbial transformations might contribute to its degradation in the ocean. Marine bacterioplankton, such as *Alcaligenes faecalis*, degrade DMSP by demethylation to methylmercaptoacetone²² or by lyase-mediated cleavage to DMS and acrylate^{9,23}. We tested the capability of common marine bacteria to degrade DMSOP through a similar pathway. After addition of $^{13}\text{C}_2$ -labelled DMSOP to an *A. faecalis* culture, DMSO with a $>99\%$ degree of $^{13}\text{C}_2$ -labelling was detected within 24 h, indicating that DMSOP was the exclusive source for DMSO production in this bacterium (Extended Data Fig. 5). Quantification of DMSO after

incubated with 1 μM DMSOP. *P* values directly over bars indicate significant difference from $t = 10$ min of the same treatment, *P* values over brackets indicate significant difference between treatment and the control without DMSOP addition. $n = 4$ independent biological replicates for 24 h, $n = 3$ for 10 min and 5 h, for statistical details see Methods.

reduction to DMS indicated that all tested bacteria (*Sulfitobacter* sp., *Ruegeria pomeroyi*, *A. faecalis* and *Halomonas* sp.) produced DMSO from DMSOP with different efficacies²¹ (Fig. 3 and Extended Data Fig. 5). By analogy with DMSP lyase-mediated cleavage, abstraction of the DMSOP alpha proton, followed by release of DMSO and acrylate is a plausible mechanism^{24,25}, supported by the observed DMSO release upon base treatment of DMSOP that occurs similarly to base-mediated DMS release from DMSP (Extended Data Fig. 4). *A. faecalis*, a bacterium with the well-identified DMSP lyase DddY and a mutant in which this enzyme was knocked out^{24,26} both showed similar DMSO production, suggesting that this DMSP lyase was not involved in this DMSOP transformation (Extended Data Fig. 5). It remains to be verified whether other reported DMSP lyases or a specific DMSOP lyase catalyse this transformation.

Our results demonstrate that the ubiquitous zwitterionic metabolite, DMSOP, contributes to the marine DMSO pool and may partly account for DMSO in marine algae¹³. In light of our findings, a functional role of DMSP as an oxygen acceptor is probable and could explain numerous observations of DMSP regulation under oxidative stress. Algal and bacterial DMSOP biosynthesis and its bacterial degradation to DMSO represent a previously undescribed pathway for DMSO production, extending our current paradigm of the marine sulfur cycle beyond the established biotic and photochemical pathways through DMS oxidation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0675-0>.

Received: 29 September 2017; Accepted: 29 August 2018;
Published online 31 October 2018.

1. Ksionzek, K. B. et al. Dissolved organic sulfur in the ocean: biogeochemistry of a petagram inventory. *Science* **354**, 456–459 (2016).
2. Sievert, S. M., Kiene, R. P. & Schulz-Vogt, H. N. The sulfur cycle. *Oceanography (Wash. DC)* **20**, 117–123 (2007).
3. Lana, A. et al. An updated climatology of surface dimethylsulfide concentrations and emission fluxes in the global ocean. *Glob. Biogeochem. Cycles* **25**, GB1004 (2011).
4. Charlson, R. J., Lovelock, J. E., Andreae, M. O. & Warren, S. G. Oceanic phytoplankton, atmospheric sulfur, cloud albedo and climate. *Nature* **326**, 655–661 (1987).

5. Lee, P. A. & de Mora, S. J. A review of dimethylsulfoxide in aquatic environments. *Atmosphere-ocean* **37**, 439–456 (1999).

6. Sunda, W., Kieber, D. J., Kiene, R. P. & Huntsman, S. An antioxidant function for DMSP and DMS in marine algae. *Nature* **418**, 317–320 (2002).

7. Kiene, R. P., Linn, L. J. & Bruton, J. A. New and important roles for DMSP in marine microbial communities. *J. Sea Res.* **43**, 209–224 (2000).

8. Alcolombrí, U. et al. Identification of the algal dimethyl sulfide-releasing enzyme: a missing link in the marine sulfur cycle. *Science* **348**, 1466–1469 (2015).

9. Todd, J. D. et al. Structural and regulatory genes required to make the gas dimethyl sulfide in bacteria. *Science* **315**, 666–669 (2007).

10. Yoch, D. C. Dimethylsulfoniopropionate: its sources, role in the marine food web, and biological degradation to dimethylsulfide. *Appl. Environ. Microbiol.* **68**, 5804–5815 (2002).

11. Asher, E. C., Dacey, J. W. H., Stukel, M., Long, M. C. & Tortell, P. D. Processes driving seasonal variability in DMS, DMSP, and DMSO concentrations and turnover in coastal Antarctic waters. *Limnol. Oceanogr.* **62**, 104–124 (2017).

12. Hatton, A. D., Shenoy, D. M., Hart, M. C., Mogg, A. & Green, D. H. Metabolism of DMSP, DMS and DMSO by the cultivable bacterial community associated with the DMSP-producing dinoflagellate *Scrippsiella trochoidea*. *Biogeochemistry* **110**, 131–146 (2012).

13. Lee, P. A. & de Mora, S. J. Intracellular dimethylsulfoxide (DMSO) in unicellular marine algae: speculations on its origin and possible biological role. *J. Phycol.* **35**, 8–18 (1999).

14. Lidbury, I. et al. A mechanism for bacterial transformation of dimethylsulfide to dimethylsulfoxide: a missing link in the marine organic sulfur cycle. *Environ. Microbiol.* **18**, 2754–2766 (2016).

15. Boden, R., Murrell, J. C. & Schäfer, H. Dimethylsulfide is an energy source for the heterotrophic marine bacterium *Sagittula stellata*. *FEMS Microbiol. Lett.* **322**, 188–193 (2011).

16. Gebser, B. & Pohnert, G. Synchronized regulation of different zwitterionic metabolites in the osmoadaptation of phytoplankton. *Mar. Drugs* **11**, 2168–2182 (2013).

17. Spielmeyer, A. & Pohnert, G. Direct quantification of dimethylsulfoniopropionate (DMSP) with hydrophilic interaction liquid chromatography/mass spectrometry. *J. Chromatogr. B* **878**, 3238–3242 (2010).

18. Carlé, J. S. & Christensen, C. Dogger Bank Itch. 4. An eczema-causing sulfoxonium ion from the marine animal, *Alcyonium gelatinosum* [Bryozoa]. *Toxicicon* **20**, 307–310 (1982).

19. Warabi, K. et al. Dogger Bank Itch revisited: isolation of (2-hydroxyethyl) dimethylsulfoxonium chloride as a cytotoxic constituent from the marine sponge *Theonella* aff. *mirabilis*. *Comp. Biochem. Physiol. B* **128**, 27–30 (2001).

20. Curson, A. R. J. et al. Dimethylsulfoniopropionate biosynthesis in marine bacteria and identification of the key gene in this process. *Nat. Microbiol.* **2**, 17009 (2017).

21. Kinsey, J. D., Kieber, D. J. & Neale, P. J. Effects of iron limitation and UV radiation on *Phaeocystis antarctica* growth and dimethylsulfoniopropionate, dimethylsulfoxide and acrylate concentrations. *Environ. Chem.* **13**, 195–211 (2016).

22. Howard, E. C. et al. Bacterial taxa that limit sulfur flux from the ocean. *Science* **314**, 649–652 (2006).

23. Curson, A. R. J., Todd, J. D., Sullivan, M. J. & Johnston, A. W. B. Catabolism of dimethylsulfoniopropionate: microorganisms, enzymes and genes. *Nat. Rev. Microbiol.* **9**, 849–859 (2011).

24. Ansede, J. H., Pellechia, P. J. & Yoch, D. C. Metabolism of acrylate to beta-hydroxypropionate and its role in dimethylsulfoniopropionate lyase induction by a salt marsh sediment bacterium, *Alcaligenes faecalis* M3A. *Appl. Environ. Microbiol.* **65**, 5075–5081 (1999).

25. Kirkwood, M., Le Brun, N. E., Todd, J. D. & Johnston, A. W. B. The *dddP* gene of *Roseovarius nubinhibens* encodes a novel lyase that cleaves dimethylsulfoniopropionate into acrylate plus dimethyl sulfide. *Microbiology* **156**, 1900–1906 (2010).

26. Curson, A. R. J., Sullivan, M. J., Todd, J. D. & Johnston, A. W. B. DddY, a periplasmic dimethylsulfoniopropionate lyase found in taxonomically diverse species of Proteobacteria. *ISME J.* **5**, 1191–1200 (2011).

Acknowledgements We thank A. Curson for the provision of the *A. faecalis* mutant; R. Kiene and A. Rellinger, M. Galí, M. Vila, L. Viure and E. Berdalet for collection of field samples and chlorophyll a analyses during field campaigns in the northeast Pacific, Arctic and Mediterranean Sea. We acknowledge the funding by the German Research Foundation (CRC1127 ChemBioSys to G.P. and N.M.), the Max Planck Society (IMPRS BGC) and the National Science Foundation (OCE-1756907 to D.J.K.). This study was co-financed by the State of Thuringia/Thüringer Aufbaubank (2015 FGI 0021) with means of the EU in the framework of the EFRE programme.

Reviewer information *Nature* thanks G. Siuzdak and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions G.P., K.T., B.G. and D.J.K. designed the research. B.G. identified DMSOP signals, performed the synthesis and the initial screening of the metabolite. K.T. performed DMSOP quantification, experiments on the biosynthesis and transformation in algae and bacteria. N.M. carried out experiments on DMSOP production and transformation in algae and performed several analytical measurements. The *I. galbana* growth experiment and DMSO quantification was performed by L.C. D.J.K. was responsible for field sampling and sample work-up. K.T. and N.M. performed the statistical evaluation of the data. G.P. and D.K. were the principal investigators for their respective research teams. K.T. and G.P. wrote the main drafts of the manuscript. All authors discussed the results and provided feedback and revisions to the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0675-0>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0675-0>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to D.J.K. or G.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Synthesis of DMSOP. The synthesis of DMSOP was carried out according to previously published studies^{27,28}. To a stirred solution of 100 mg 3-dimethylsulfonylpropionate (DMSP) (synthesized as previously described²⁹) in 0.5 ml deionized water, 0.24 ml of an aqueous 0.24 M RuCl₃ hydrate solution (Carl Roth) was added at room temperature. A 12% sodium hypochlorite solution (12% Cl, Carl Roth) was added dropwise at room temperature to the dark solution until the colour changed to a yellowish green. When the solution turned brown after stirring for a few minutes more sodium hypochlorite was added. The pH was adjusted to 5–6 with a 1 M HCl solution (37%, Carl Roth) during the reaction. When the solution did not embrown further, the water was removed in a rotary evaporator and the resulting white solid was dissolved at room temperature in a minimum amount of methanol. Diethylether (Et₂O) was added dropwise until a precipitate formed. The precipitate settled within 30 min, and additional Et₂O was added. This procedure was repeated until no further precipitate formed. The precipitate was filtered off and dried on the filter. Owing to salt residues in the product, elemental analysis—based on sulfur content in the final product relative to theoretical sulfur content of pure DMSOP—was used to determine the degree of purity.

¹H NMR (600 MHz, D₂O) δ p.p.m.: 3.21 (2H, t, *J*=6.88 Hz, H(C2)), 3.89 (6H, s, H(C4)), 4.33 (2H, t, *J*=6.88 Hz, H(C3)); ¹³C NMR (200 MHz, D₂O) δ p.p.m.: 25.49 (C2), 38.68 (C4), 48.48 (C3), 172.78 (C1). Numbering of carbons and heteronuclear multiple-bond coherence correlations are shown in Extended Data Fig. 2. ESI-MS (positive) *m/z* 151.56 [M + H]⁺; ESI-MS/MS (parent ion *m/z* 151, collision energy 15 eV): *m/z* 151.56 [M + H]⁺, 79.30 [M – C₃H₅O₂ + H]⁺, 73.29 [C₃H₄O₂ + H]⁺; elemental analysis: calculated C 32.2%, H 5.9%, S 17.2%, Cl 19.0%; found C 25.5%, H 4.7%, S 13.4%, Cl 26.7%; degree of purity 77.8%.

Synthesis of ¹³C₂-DMSOP was done as described above using ¹³C₂-DMSP as starting material. This was synthesized using ¹³C₂-DMS according to a previous study²⁹.

Cultivation of phytoplankton. Cultures were obtained from the Provasoli-Guillard National Center for Marine Algae and Microbiota, East Boothbay (CCMP strains), the Roscoff Culture Collection, Roscoff (RCC strains), the UTEX Algae Express, Austin (UTX strains), and the Culture Collection of Algae and Protozoa, Oban (SCCAP strains). Axenic *I. galbana* (CCMP 1323, https://ncma.bigelow.org/ccmp1323?__SID=U#.W8HwWvIRcf) batch cultures were grown in a modified Guillard f/2 medium without silica in 2.8-l Fernbach flasks. The modified f/2–Si medium consisted of 1 l of autoclaved 0.2-μm-filtered Sargasso seawater (salinity 34.9 p.p.t.) enriched with 160 μM NaNO₃, 10 μM NaH₂PO₄, 1.0 μM Fe, 11.7 μM EDTA, 39.9 nM Cu, 26.0 nM Mo, 76.5 nM Zn, 42.0 nM Co, 910 nM Mn, 296 nM vitamin B₁, 2.05 nM biotin and 0.369 nM vitamin B₁₂.

I. galbana cultures were grown under batch conditions with cool white-fluorescent lighting (92.7 μmol photons m⁻² s⁻¹ between 400 and 700 nm) with a 14:10 h day:night cycle in an incubator (model I-36 LLVL, Percival Scientific). The temperature was maintained at 23.0 ± 0.1 °C. Daily sampling started at 10:00 local time. Axenicity was periodically determined by DAPI staining followed by epifluorescence microscopy counting²¹.

For DMSOP screening, *S. costatum* RCC75, *I. galbana*, *Chaetoceros compressum* CCMP168, *Chaetoceros didymus* CH5, *Entomoneis paludosa*, *Nitzschia cf. pellucida* DCG0303, *Navicula* sp. I15, *Phaeodactylum tricornutum* CCMP2561, SCCAP K-128 and UTX646, axenic *P. parvum* CCAP 946/6, *Stephanopyxis turris*, *Thalassiosira pseudonana* CCMP1335, *Thalassiosira rotula* RCC841, RCC776 and CCMP1018, *Thalassiosira weissflogii* RCC76 and *Rhodomonas* sp. were cultivated in an artificial seawater medium³⁰. *Phaeocystis pouchetii* AJ01, *Amphidinium carterae* SCCAP K-0406 and *P. minimum* were cultivated in f/2 medium³¹. No silicate was added to the medium used to cultivate *P. minimum*. *Coscinodiscus wailesii* CCMP2513, *Lingulodinium polyedrum* CCAP1221/2 and *Symbiodinium microadriaticum* CCMP2464 were cultivated in L1 medium³²; no silicate was added to the *S. microadriaticum* L1 medium. The medium for *E. huxleyi* was prepared according to a previously published study³³. Cultivation was done from stock cultures by 20-fold dilution of a cell suspension in tissue-culture flasks. Cultures were grown in a 14:10 h light:dark cycle with light provided by osram biolux lamps (40 μmol m⁻² s⁻¹ between 400 and 700 nm) at 12 °C, except for *P. pouchetii* which was cultivated at 5 °C. Cultures were grown to the exponential phase and then divided into four aliquots of equal volume. These aliquots were 20-fold diluted with fresh medium and cultivated again to the exponential phase before being used for quantitative analysis as described below.

For all cultures except for *I. galbana*, cell counts were determined in a Fuchs-Rosenthal haemocytometer using a Leica DM2000 upright microscope with phase contrast. Cell volumes for *P. minimum* and *E. huxleyi* were obtained from a previous study¹⁶, whereas other cell volumes were calculated according to a previously published method³⁴. Cell counts and cell volumes for *I. galbana* cultures were determined by adding 200 μl of an unfiltered sample to 10 ml of 0.2-μm-filtered electrolyte diluent (1% sodium chloride in 50 mM phosphate buffer, pH 7.4).

Samples were analysed with a Beckman–Coulter Z2 Particle Counter and Size Analyzer fitted with a 100-μm aperture.

The photosynthetic efficiency of photosystem II (*F_v/F_m*) was determined during the *I. galbana* growth experiment using a model Water-PAM, pulse-amplitude-modulated fluorometer (Walz). To determine *F_v/F_m*, triplicate 3-ml aliquots of unfiltered culture samples were dark-adapted at room temperature for 30 min. The fluorometer was blanked with 0.2-μm-filtered Sargasso seawater. After 30 min, a saturating pulse (~3,230 μmol m⁻² s⁻¹, 0.6 s) was applied to each culture sample for a total of six to eight measurements. Sample dilutions were performed as needed with 0.2-μm-filtered Sargasso seawater. Gain settings were 2–3 for the photomultiplier and 1 for signal output, except for early in the growth curve when the photomultiplier gain was set at 6 and signal output gain was set at 5.

Cultivation of bacteria. *Halomonas* sp. HTNK-1, *A. faecalis* M3A and the *dddY* knockout mutant of *A. faecalis* M3A (obtained from A. Curson, University of East Anglia²⁰) were grown in M9 minimal medium (Sigma-Aldrich). *R. pomeroyi* DSS-3 and *Sulfobacter* sp. EE-36 were grown in a marine basal medium. The cultures were grown under gentle shaking at 28 °C with addition of 10 mM sodium succinate as the carbon source. For the incubation experiment, experimental cultures were prepared in four replicates for each sampling point from the stock culture by a 20-fold dilution of an aliquot of cell suspension in tissue-culture flasks and grown to exponential phase. *P. bermudensis* DSM 15984 (Deutsche Sammlung von Mikroorganismen und Zellkulturen) was cultivated in Marine Broth medium (Carl Roth) and grown under gentle shaking at 28 °C.

Field samples. Unfiltered seawater samples were collected from the near surface in Niskin bottles attached to a CTD rosette. For each sample, triplicate 15-ml subsamples were collected directly from the Niskin bottle into three precleaned and baked (550 °C, 8 h) 20-ml glass scintillation vials, each with a green thermoset screw cap containing a Teflon-faced silicone insert. Samples were collected on three oceanographic cruises: the northwest Atlantic on the R/V *Endeavour*, the northeast Pacific aboard the R/V *Oceanus* and in the Arctic aboard the Canadian research icebreaker CCGS *Amundsen*. The Mediterranean Sea samples were collected in 250-ml pre-cleaned polyethylene bottles (prerinsed with 5% HCl followed by high-purity laboratory water) just below the sea surface; one sample was collected offshore just beyond the breaking waves and one sample was collected nearshore in the wave breaking zone. A map of the sampling locations is shown in Extended Data Fig. 6.

Each sample vial was microwaved to boiling (approximately 12 s) with the cap loose. Once the sample cooled to room temperature, it was bubbled with ultra-high purity He (99.9995%) for 10 min to quantitatively remove DMS (verified by testing for residual DMS by resparging the same sample), and then 150 μl of Ultrex concentrated HCl (Baker) was added to each sample to preserve DMS and DMSOP in their protonated forms followed by storage in the dark at room temperature until analysis.

For chlorophyll *a* samples, 5–50 ml of unfiltered seawater was filtered with a low vacuum (approximately 130 mbar) through a prebaked (550 °C, 8 h) GF/C filter (Whatman), and the folded filter was placed into a 10-ml borosilicate test tube that was stored at –20 °C until analysis. Unless otherwise noted, triplicate samples were filtered. Chlorophyll *a* samples were analysed by adding 5 ml of 90% acetone (10% water) to each test tube. Samples were vortexed and then allowed to incubate overnight at –20 °C. The chlorophyll fluorescence was then measured with a TD-700 fluorometer³⁵.

Extraction and sample preparation of phytoplankton cellular DMSOP. For all cultures except *I. galbana*—to screen for the presence of particulate DMSOP—algal cultures were filtered under reduced pressure (GF/C grade microfibre filter; GE healthcare) at 400 mbar. Particulate DMSOP in *I. galbana* samples were collected by small-volume gravity filtration³⁶. The filters were immediately transferred to 4-ml glass vials containing 1 ml of methanol and vortexed. Extracts were stored at –20 °C. To prepare a sample for liquid chromatography–mass spectrometry (LC-MS) analysis, 50 μl of the extract was diluted with 100 μl of a mixture of acetonitrile and water (9:1 v/v). For ultra performance liquid chromatography–mass spectrometry (UPLC-MS) analysis, 10 μl of an aqueous solution of the internal standard D₆-dimethylsulfonyoacetate (D₆-DMSA) was added to the extract before injection. The D₆-DMSA was synthesized according to previously published studies^{16,37}. After centrifugation (5 min, 4,500g) the supernatant was submitted to LC-MS analysis.

Extraction and sample preparation of dissolved DMSOP. To quantify dissolved DMSOP, a dense *P. parvum* culture was divided into four aliquots of equal volume and 20-fold diluted with fresh medium. On day 1, 5, 7 and 11, 1 ml of culture was centrifuged in an Eppendorf tube for 5 min at 100g. The supernatant was transferred to a 1.5-ml glass vial and 5 μl was directly submitted to ultra-high-pressure liquid chromatography/high-resolution mass spectrometry (UHPLC/HRMS) for analysis.

Extraction and sample preparation of bacterial DMSOP. Aliquots of the bacterial cultures (100 μl) were centrifuged for 5 min at 16,100g and the supernatant was removed by pipetting. The pellets were taken up in 100 μl of a mixture of acetonitrile and water (9:1 v/v) and samples were frozen at –20 °C and stored overnight.

After thawing the samples, cells were disrupted by sonication using ten pulses in a Bandelin Sonoplus ultrasound homogenizer (Bandelin). The samples were again centrifuged for 5 min at 16,100g and 5 μ l of the supernatant was directly submitted to UHPLC/HRMS for analysis.

Extraction and sample preparation of field samples. For determination of DMSOP in field samples, a 3-ml sample was freeze-dried and redissolved in 500 μ l acetonitrile. Owing to the high salt content of the sample, a precipitate remained that settled. The supernatant was transferred to a 1.5-ml glass vial and dried in a gentle nitrogen stream at 30 °C and resolved in 300 μ l of a mixture of acetonitrile and water (9:1 v/v). After centrifugation (5 min, 4,500g), the supernatant was stored at –80 °C until UPLC/MS measurement.

UPLC/MS analysis. Analytical separation and quantification of DMSOP in the algal extracts for results shown in Fig. 2, Table 1 and Extended Data Table 1 were performed using an Acquity UPLC (Waters) equipped with a SeQuant ZIC-HILIC column (5 μ m, 2.1 mm \times 150 mm, SeQuant). Quantification followed a previously reported protocol with modifications as follows³⁸. The eluent consisted of high-purity water with 2% acetonitrile and 0.1% formic acid (solvent A) and 90% acetonitrile with 10% 5 mmol l^{–1} aqueous ammonium acetate (solvent B). The flow rate was set to 0.60 ml min^{–1}. A linear gradient was used for separation with 100% solvent B (1 min), 20% B (6.5 min), 100% B (7.1 min) and 100% B (10 min). The column was kept at 25 °C. A Q-ToF micro mass spectrometer (Waters Micromass) with electrospray ionization in positive mode was used as the mass analyser. The sample cone was set to 18 V, the extraction cone to 1 V, the sheath gas was operated at 20 l h^{–1} and the desolvation gas at 450 l h^{–1}. MS/MS for fragmentation of DMSOP was accomplished with a collision energy of 15 eV. Calibration curve: area [DMSOP] = 123c [DMSOP in μ M] with $r = 0.9983$, limit of detection (LOD) = 0.08 μ M, limit of quantification (LOQ) = 0.1 μ M. Data analyses were done using the software MassLynx version 4.1.

UHPLC/high-resolution MS analysis. All other LC/MS results were obtained on a Dionex Ultimate 3000 system (Thermo Scientific) coupled to a Q Exactive Plus Orbitrap mass spectrometer (Thermo Scientific). Electrospray ionization was performed in positive mode ionization with the following parameters: capillary temperature 380 °C, spray voltage 3,000 V, sheath gas flow 60 arbitrary units and aux gas flow 20 arbitrary units. The LC separation column and the solvent gradient were identical to that described in 'UPLC/MS analysis'; the injection volume was 5 μ l.

Calibration curves for DMSP and DMSOP were recorded in triplicate using synthetic standards prepared as described above and in a previous study²⁹. For DMSOP, the LOD was 0.01 nM, the LOQ was 0.1 nM and the linear range was between 0.1 and 1,000 nM. Calibration curve: area [DMSOP] = 418,370c [DMSOP in nM] with $r = 0.9998$. For DMSP, the calibration curve was: area [DMSP] = 470,540c [DMSP in nM] with $r = 0.9999$. MS/MS for fragmentation of DMSOP was accomplished with a normalized collision energy of 35. Data analyses were done using the software Thermo Xcalibur version 3.0.63.

DMSO quantification using purge and trap gas chromatography/flame photometric detection. Analyses of samples to quantify DMSO were done according to a previous study²¹. In brief, 3 ml of unfiltered culture samples were pipetted into 4-ml glass vials (see 'DMSOP transformation' for details) and stored frozen until analysis. For analysis, samples were first tested to see whether they contained DMS. Because no DMS was detected in the samples, they were not bubbled with ultra-high purity He to remove DMS before analysis. The total DMSO concentration in unfiltered culture samples or medium controls was measured after reduction to DMS by TiCl₃, as previously described²¹. For each sample, a 1-ml aliquot was mixed with 200 μ l TiCl₃ reagent (20% w/v in 2 M HCl, EMD Chemicals) in a 14-ml serum vial that was crimp-sealed with a Teflon-lined butyl rubber stopper and an aluminium crimp cap. The DMSO samples were allowed to react for 1 h at 55 °C, then allowed to cool down to room temperature for analysis.

Reacted vials containing DMS were sparged with ultra-high purity He for 3 min to transfer the DMS from the vials onto liquid-nitrogen-cooled Teflon wool using a custom-made cryogenic purge-and-trap system. Hot water (approximately 90 °C) was used to desorb the DMS from the Teflon wool and inject the sample into Shimadzu model GC-14A gas chromatograph equipped with a Chromosil 330 column (2.4 m long \times 3.2 mm inner diameter, Supelco). The sulfur was detected with a sulfur-selective flame photometric detector. The column temperature was set isothermally at 60 °C. Both the injection port and detector temperature were set at 225 °C. Authentic DMSP and DMSO standards were prepared in the same manner as the samples. The LOD of the method is 0.2 pmol S for a 1-ml aqueous sparged sample, with a signal-to-noise ratio of two.

Confirmation of DMSOP in the algal extract. A *P. parvum* methanolic extract from a stationary growth-phase culture was used to determine whether the signal of the unknown metabolite in the extract co-eluted with an authentic DMSOP standard that was added to the extract before injection into the UPLC. As a control, 50 μ l of the extract with no DMSOP standard was diluted with 100 μ l of a mixture of acetonitrile and water (9:1 v/v). After centrifugation (5 min, 4,500g),

the supernatant was injected into the UPLC. In a separate analysis, an aliquot of this *P. parvum* extract was mixed with 10 μ l of a 10 μ M DMSOP standard solution, and then prepared for analysis in the same way as the control. Comparison of the peaks of mass trace m/z = 151 for the two injections showed an increased area at a retention time of $t_R = 4.2$ min corresponding to the DMSOP-containing extract.

¹³C₂-DMSP transformation. *P. bermudensis* cultures (6.5 ml, optical density (OD₆₀₀) = 1.97 \pm 0.05, protein content = 99 \pm 1.3 μ g ml^{–1}, $n = 3$) were concentrated by centrifugation to 1 ml before addition of 10 μ l of ¹³C₂-DMSP (10 mM in H₂O). Samples were maintained under shaking at 28 °C for 18 h. Aliquots (100 μ l) of the cultures were centrifuged and the pellet was treated as previously described for the extraction and sample preparation of bacterial DMSOP.

DMSOP transformation. Prior to incubation, aliquots of the bacterial cultures (10–15 ml) were washed three times by centrifugation (15 min, 4,500g) and subsequently resuspended in 10 ml of a succinate-free medium to remove excess organic carbon. For incubation experiments, all bacterial cultures were diluted with succinate-free medium to an OD = 0.10–0.12. Culture samples (3 ml each) were transferred into 4-ml screw cap vials with PTFE/silicone septa. After addition of either an aqueous DMSOP solution (0.65 mM) to a final concentration of 1 μ M or the same amount of water (controls), the vials were sealed, vortexed and placed on a shaker at 28 °C. Samples and controls were prepared for each culture in four replicates. Samples were taken directly after substrate addition (10 min), and after 5 and 24 h. The vials were frozen at –20 °C until DMSO quantification. As controls, marine basal medium (MBM) and M9 medium with added DMSOP at a final concentration of 1 μ M were prepared in four replicates. Incubation conditions and sampling times were done as described above.

Gas chromatography/high-resolution MS measurement of ¹³C₂-DMSO. To determine whether DMSOP was a DMSO precursor, we developed a method to detect DMSO using solid-phase microextraction (SPME) in combination with gas chromatography (GC)/high-resolution (HR) MS. DMSO was extracted from set ups as described above for DMSOP transformation in 4-ml glass vials sealed with PTFE septa. Extraction was achieved with a SPME fibre (100 μ m PDMS, Supelco). Prior to extraction, the SPME fibre was conditioned for 15 min at 250 °C. To apply the fibre to the sample vial, a hole was pierced in the septum and the needle of the SPME holder was inserted into the vial. By immersion of the fibre into the constantly stirred solution the analyte was allowed to adsorb onto the fibre for 15 min at room temperature. Subsequently, the fibre was inserted into the injection port of the GC. DMSO was desorbed into the PTV injector at 300 °C for 5 min in a gas chromatograph (TRACE 1310, Thermo Scientific) that was fitted with a 60 m \times 0.25 mm, 1 μ m film ZB-1MS capillary column (Phenomenex) and a hybrid quadrupole-orbitrap mass spectrometer (Q Exactive, Thermo Scientific). Ultra-high purity helium was used as carrier gas at a flow rate of 1.2 ml min^{–1}. The oven temperature was held for 1 min at 40 °C and subsequently increased to 150 °C (15 °C min^{–1}) and again held for 3.5 min. The transfer line and ion source were both set to 300 °C. Mass measurements were performed in electron ionization-positive mode. A mass range from 45 to 200 m/z was recorded. The ionization energy was 70 eV and scan time 0.25 s. Data analyses were performed with the Thermo Xcalibur software version 3.0.63.

DMSOP base lability. A 0.5 M NaOH solution (2.5 μ l) was added to 1 ml of an aqueous DMSOP solution in water (500 μ M). A DMSOP solution without addition of NaOH served as a control. Samples were prepared in triplicate. To determine DMSO, samples (50 μ l) were taken immediately after the addition of NaOH (0 min), and after a reaction time of 2.5, 5.3 and 23 h at room temperature. DMSO was detected by UHPLC/HRMS using a Rezex ROA-Organic Acid (8%) column (8 μ m, 4.6 \times 150 mm, Phenomenex). Separations were carried out isocratically at 90% 0.0025 M trifluoroacetic acid (solvent A) and 10% acetonitrile (solvent B) for 12 min. The flow rate was set to 0.40 ml min^{–1}. DMSOP was quantified as described above.

Statistical analysis. Data are given as mean \pm s.d., the number of replicates (n) is listed. For comparison of two groups an unpaired two-tailed *t*-test was used. As prerequisites, normal distribution (Shapiro–Wilk test) and equal variance were tested. If at least one of those prerequisites was not met ($P \geq 0.05$), a Mann–Whitney *U*-test was performed. For comparison of multiple time points, a one-way analysis of variance (ANOVA) was used. If prerequisites were not met, a Kruskal–Wallis one-way ANOVA on ranks was performed. If samples were drawn repeatedly from the same vessel, a one-way repeated-measures ANOVA was used. All ANOVAs were followed by a Tukey post hoc test for multiple pairwise comparisons if there was a significant difference in the dataset. All statistical analyses were performed with a 95% confidence interval using Sigma-Plot version 11.0. $P > 0.05$ is considered not significantly different. For results in Fig. 3b, no equal variance was observed within the treatment 'control' and Kruskal–Wallis one-way ANOVA on ranks with Tukey post hoc test for different time points was conducted. Within the treatment '+DMSOP' a one-way ANOVA with Tukey post hoc test for different time points was conducted. Within time points 10 min and 24 h, unpaired two-tailed Student's *t*-tests between 'control' and '+DMSOP' were performed. Within

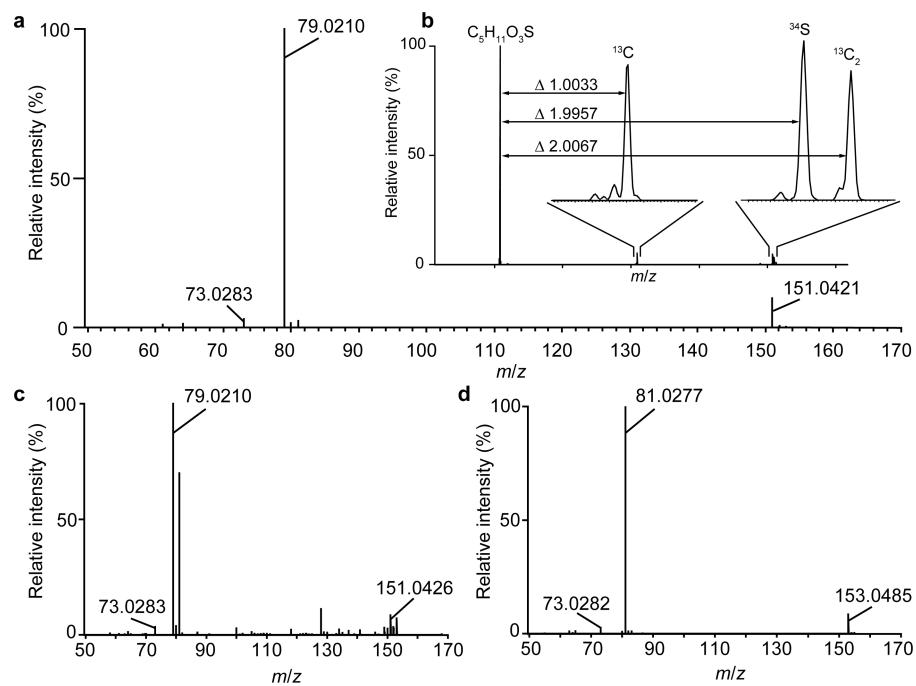
time point 5 h, a normal distribution was not observed and therefore a Mann-Whitney *U*-test was conducted to compare between the control and treatment. For results in Fig. 3c, a one-way ANOVA with Tukey post hoc test for different time points was conducted within the treatment 'control' and within the treatment '+DMSOP'. Within time points 10 min and 5 h, unpaired two-tailed Student's *t*-test between 'control' and '+DMSOP' were conducted. For the 24-h time point, no equal variance was found and a Mann-Whitney *U*-test was performed. The loss of a medium control sample during transport led to the exclusion of one replicate of the treatment '+DMSOP' ($t=10$ min) from the analysis in Fig. 3b, c. A contaminated medium control sample led to exclusion of a replicate of the treatment '+DMSOP' ($t=5$ h) from the analysis in Fig. 3b, c.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

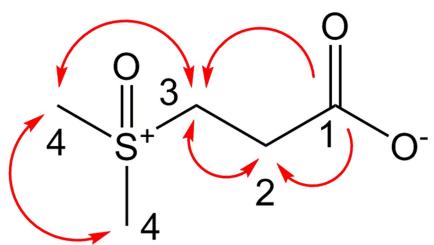
The datasets generated and analysed during the current study are available from the corresponding authors upon reasonable request.

27. Forrester, J., Jones, R. V. H., Preston, P. N. & Simpson, E. S. C. Generation of trimethylsulfonium cation from dimethyl sulfoxide and dimethyl sulfate: implications for the synthesis of epoxides from aldehydes and ketones. *J. Chem. Soc. Perkin Trans. I* **0**, 2289–2291 (1995).
28. Ayres, D. C. & Hossain, A. M. M. Oxidation of aromatic substrates. Part II. The action of ruthenium tetroxide on some derivatives of napthalene and its monoaza-analogues. *J. Chem. Soc. Perkin Trans. I* **0**, 707–710 (1975).
29. Chambers, S. T., Kunin, C. M., Miller, D. & Hamada, A. Dimethylthetin can substitute for glycine betaine as an osmoprotectant molecule for *Escherichia coli*. *J. Bacteriol.* **169**, 4845–4847 (1987).
30. Maier, I. & Calenberg, M. Effect of extracellular Ca^{2+} and Ca^{2+} -antagonists on the movement and chemorientation of male gametes of *Ectocarpus siliculosus* (Phaeophyceae). *Bot. Acta* **107**, 451–460 (1994).
31. Guillard, R. R. & Ryther, J. H. Studies of marine planktonic diatoms. I. *Cyclotella nana* Hustedt, and *Detonula confervacea* (Cleve) Gran. *Can. J. Microbiol.* **8**, 229–239 (1962).
32. Guillard, R. R. L. & Hargraves, P. E. *Stichochrysis immobilis* is a diatom, not a chrysophyte. *Phycologia* **32**, 234–236 (1993).
33. Spielmeyer, A., Gebser, B. & Pohnert, G. Investigations of the uptake of dimethylsulfoniopropionate by phytoplankton. *ChemBioChem* **12**, 2276–2279 (2011).
34. Verity, P. G. et al. Relationships between cell volume and the carbon and nitrogen content of marine photosynthetic nanoplankton. *Limnol. Oceanogr.* **37**, 1434–1446 (1992).
35. Welschmeyer, N. A. Fluorimetric analysis of chlorophyll *a* in the presence of chlorophyll *b* and pheophytins. *Limnol. Oceanogr.* **39**, 1985–1992 (1994).
36. Kiene, R. P. & Slezak, D. Low dissolved DMSP concentrations in seawater revealed by small-volume gravity filtration and dialysis sampling. *Limnol. Oceanogr. Methods* **4**, 80–95 (2006).
37. Howard, A. G. & Russell, D. W. Borohydride-coupled HPLC-FPD instrumentation and its use in the determination of dimethylsulfonium compounds. *Anal. Chem.* **69**, 2882–2887 (1997).
38. Spielmeyer, A., Gebser, B. & Pohnert, G. Dimethylsulfide sources from microalgae: improvement and application of a derivatization-based method for the determination of dimethylsulfoniopropionate and other zwitterionic osmolytes in phytoplankton. *Mar. Chem.* **124**, 48–56 (2011).

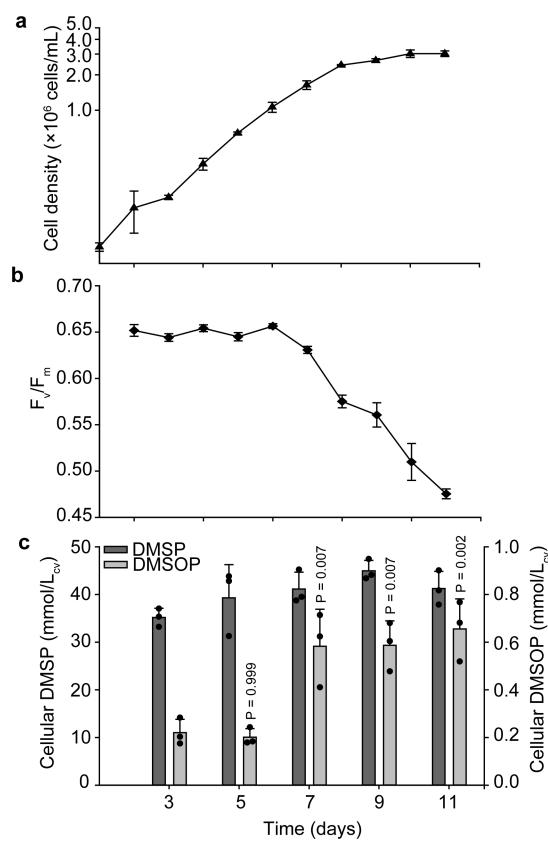


Extended Data Fig. 1 | DMSOP mass spectra. The HRMS/MS spectra of natural occurring DMSOP and the authentic standard (normalized collision energy of 35) are shown. **a**, DMSOP standard, molecular ion m/z 151.0421, fragments $[C_2H_7O_2S]^+$ m/z 79.0210 and $[C_3H_5O_2]^+$ m/z 73.0283. **b**, Isotopic pattern of the molecular ion m/z 151.0421 with the

calculated formula $C_5H_{11}O_3S$ and isotopic fine structure of $[M + 1]$ and $[M + 2]$. **c**, DMSOP from a *P. parvum* extract with added $^{13}C_2$ -DMSOP. **d**, $^{13}C_2$ -DMSOP, molecular ion m/z 153.0485, fragments $[^{13}C_2H_7O_2S]^+$ m/z 81.0277 and $[C_3H_5O_2]^+$ m/z 73.0282.

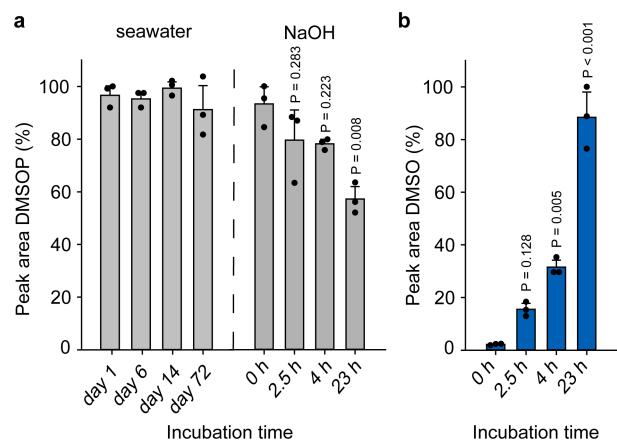


Extended Data Fig. 2 | Structure of DMSOP. Arrows show the heteronuclear multiple-bond coherence correlations. Numbers indicate carbon atom positions.

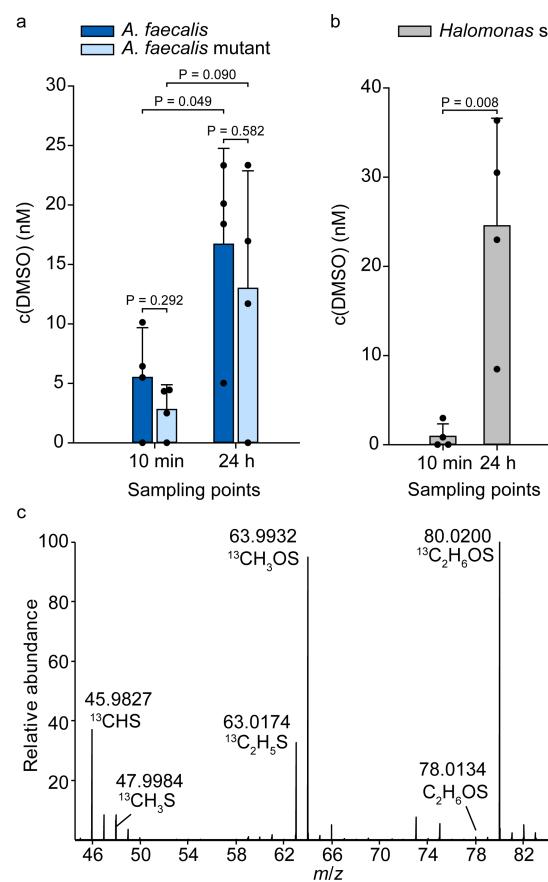


Extended Data Fig. 3 | *I. galbana* growth and cellular DMSOP.

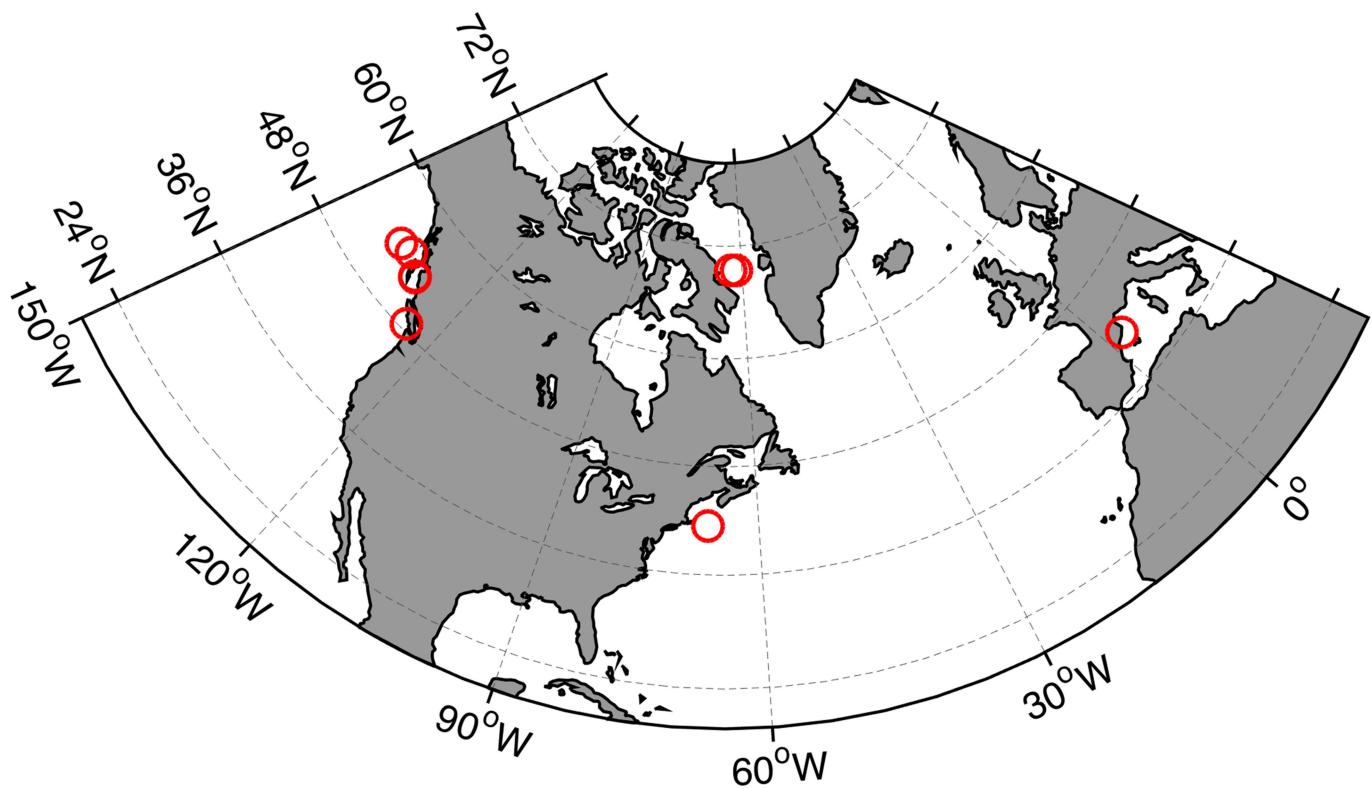
a, b, Growth (a) and photosynthetic efficiency (b) of *I. galbana* cultures. **c**, Cellular DMSP and DMSOP content. Data are mean \pm s.d. of $n = 3$ independent cultures. P values are from a one-way repeated-measures ANOVA with Tukey post hoc test. A significant difference in cellular DMSOP concentration compared to day 3 was detected from day 7 onward.



Extended Data Fig. 4 | DMSOP stability in seawater and base. **a**, DMSOP is stable over a period of 72 days in seawater (left). It degraded at room temperature under basic pH (pH = 11, monitored over 23 h; right). **b**, DMSO was released during this base treatment (integration of $m/z = 79$ in GC/MS). Data are mean \pm s.d. of $n = 3$ independent samples. P values are from one-way repeated-measures ANOVA with Tukey post hoc test compared to $t = 0$ h.


Extended Data Fig. 5 | DMSO release from DMSOP by bacteria.

a, b, DMSOP (1 μ M) is degraded by *A. faecalis*, a *dddY* knockout mutant of *A. faecalis* (**a**) and by *Halomonas* sp. (**b**). Data are mean \pm s.d. of $n = 4$ independent cultures. P values result from unpaired two-tailed Student's *t*-tests. **c**, In separate experiments, it was demonstrated that DMSOP is the exclusive source for DMSO production in *A. faecalis*. Release of labelled DMSO from ¹³C₂-DMSOP was monitored by GC/HRMS. The mass spectrum shows an average over the DMSO peak extracted from an *A. faecalis* culture that was incubated for 23 h with DMSOP. Integration of the ion traces 80.0200 (¹³C₂-DMSO) and 78.0134 (DMSO) in three independent replicates revealed a degree of labelling of $99.3 \pm 0.25\%$.



Extended Data Fig. 6 | Map of sampling sites. Sampling sites are indicated in red.

Extended Data Table 1 | Occurrence of DMSOP in different algal species

Class	Species	Strain	DMSOP
haptophyte	<i>Isochrysis galbana</i>		+
	<i>Prymnesium parvum</i> (axenic)		+
	<i>Prymnesium parvum</i>	CCAP946/6	+
	<i>Phaeocystis pouchetii</i>	AJ01	-
diatom	<i>Chaetoceros compressum</i>	CCMP168	-
	<i>Chaetoceros didymus</i>	CH5	-
	<i>Coscinodiscus wailesii</i>	CCMP2513	+
	<i>Entomoneis paludosa</i>		+
	<i>Eucampia zodiacus</i>		+
	<i>Nitzschia cf. pellucida</i>	DCG0303	-
	<i>Navicula</i> sp.	I15	-
		CCMP2561	
	<i>Phaeodactylum tricornutum</i>	SCCAP K-128	-
		UTX646	
	<i>Skeletonema costatum</i>	RCC75	+
	<i>Stephanopyxis turris</i>		-
	<i>Thalassiosira pseudonana</i>	CCMP1335	-
		RCC841	
	<i>Thalassiosira rotula</i>	RCC776	-
		CCMP1018	
	<i>Thalassiosira weissflogii</i>	RCC76	-
coccolithophore		RCC1217	
	<i>Emiliania huxleyi</i>	RCC1731	+
cryptophyceae	<i>Rhodomonas</i> sp.		-
dinoflagellate	<i>Amphidiniuum carterae</i>	SCCAP K-0406	-
		CCAP1121/2	-
	<i>Lingulodinium polyedrum</i>		
	<i>Prorocentrum minimum</i>		+
	<i>Symbiodinium microadriaticum</i>	CCMP2464	+

DMSOP measurements above the limit of detection of 0.08 µM (UPLC/MS analysis) are indicated by '+'; whereas DMSOP measurements below the limit of detection are indicated by '-'. The ratio of peak area (DMSOP)/peak area (DMSP) was >0.01% in all samples labelled with '+'. Cultures without strain denomination are from our culture stock in the laboratory of the Institute of Inorganic and Analytical Chemistry Jena (strains available upon request).

Extended Data Table 2 | DMSP_{total} and DMSOP_{total} concentrations in seawater.

Location	Date	Latitude	Longitude	Depth	Temp.	Sal.	Chl a	DMSP _t	DMSOP _t *
	(2016)	(°N)	(°W)	(m)	(°C)	(ppt)	($\mu\text{g L}^{-1}$)	(nM)	(nM)
NW Atlantic	Sept 21	41.40	67.47	5	18.5	32.5	3.14 ± 0.02	16.7 ± 1.4	$0.057 \pm 0.048^{\ddagger}$
Arctic	July 9	69.50	61.58	10	-0.7	32.8	0.47	44.8 ± 2.4	0.197 ± 0.257
	July 10	69.50	63.23	12	-1.3	32.3	0.24	37.8 ± 2.4	0.057 ± 0.043
NE Pacific	July 14	54.04	137.16	5	13.6	32.1	0.63 ± 0.01	49.3 ± 6.9	0.061 ± 0.037
	July 15	54.30	134.68	5	14.9	31.8	0.55 ± 0.01	34.1 ± 1.6	0.036 ± 0.003
	July 19	52.90	130.62	5	13.1	31.5	6.09 ± 0.12	83.1 ± 7.9	0.151 ± 0.015
	July 19	52.96	130.73	5	14.2	31.5	1.80 ± 0.01	49.7 ± 3.0	0.190 ± 0.081
	July 22	48.75	125.42	5	14.9	31.0	16.5 ± 0.57	122.0 ± 15.5	0.079 ± 0.021
Mediterranean	July 18	41.55	2.49 [‡]	surface	24.5	37.4	1.21	24.8 ± 4.5	0.073 ± 0.050
Sea	July 18	41.55	2.49 [‡]	surface	24.5	37.4	1.04	60.5 ± 6.3	0.045 ± 0.037

Data are mean \pm s.d. ($n=3$ independent samples). When no s.d. is reported $n=1$.

*LOQ = 0.1 nM, LOD = 0.01 nM (UHPLC/HRMS analysis).

[†] $n=2$.

[‡]Longitude reported in °E.

Extended Data Table 3 | Incorporation of $^{13}\text{C}_2$ -DMSP into DMSOP in *P. bermudensis*

Peak area	Peak area	Peak area	Degree of labeling [%]
$^{13}\text{C}_2$ -DMSOP	$^{13}\text{C}_1$ -DMSOP*	DMSOP	$^{13}\text{C}_2$ -DMSOP in relation to DMSOP†
<i>m/z</i> 153.0496	<i>m/z</i> 152.0455	<i>m/z</i> 151.0423	
3,140,000 \pm 640,000	4,310,000 \pm 180,000	84,700,000 \pm 4,080,000	3,68 \pm 0,59

*The area corresponds to approximately 5.1% of the unlabelled isotopologue, which is in accordance with the natural ^{13}C content of a compound with five carbon atoms (5.5%).

†Values exceed the calculated degree of labelling of the natural isotopologue of 0.26% and confirm that externally added labelled DMSP was transformed to DMSOP. Data are mean \pm s.d. of $n = 3$ independent experiments.

Metal-free ribonucleotide reduction powered by a DOPA radical in *Mycoplasma* pathogens

Vivek Srinivas^{1,6}, Hugo Lebrette^{1,6}, Daniel Lundin¹, Yuri Kutin², Margareta Sahlin¹, Michael Lerche¹, Jürgen Eirich³, Rui M. M. Branca³, Nicholas Cox⁴, Britt-Marie Sjöberg¹ & Martin Högbom^{1,5*}

Ribonucleotide reductase (RNR) catalyses the only known de novo pathway for the production of all four deoxyribonucleotides that are required for DNA synthesis^{1,2}. It is essential for all organisms that use DNA as their genetic material and is a current drug target^{3,4}. Since the discovery that iron is required for function in the aerobic, class I RNR found in all eukaryotes and many bacteria, a dinuclear metal site has been viewed as necessary to generate and stabilize the catalytic radical that is essential for RNR activity^{5–7}. Here we describe a group of RNR proteins in Mollicutes—including *Mycoplasma* pathogens—that possess a metal-independent stable radical residing on a modified tyrosyl residue. Structural, biochemical and spectroscopic characterization reveal a stable 3,4-dihydroxyphenylalanine (DOPA) radical species that directly supports ribonucleotide reduction in vitro and in vivo. This observation overturns the presumed requirement for a dinuclear metal site in aerobic ribonucleotide reductase. The metal-independent radical requires new mechanisms for radical generation and stabilization, processes that are targeted by RNR inhibitors. It is possible that this RNR variant provides an advantage under metal starvation induced by the immune system. Organisms that encode this type of RNR—some of which are developing resistance to antibiotics—are involved in diseases of the respiratory, urinary and genital tracts. Further characterization of this RNR family and its mechanism of cofactor generation will provide insight into new enzymatic chemistry and be of value in devising strategies to combat the pathogens that utilize it. We propose that this RNR subclass is denoted class Ie.

Three RNR classes have been discovered so far, and all require transition metals to function². Class III is strictly anaerobic and uses a 4Fe-4S cluster for radical generation, whereas class II is indifferent to oxygen and utilizes an adenosyl cobalamin cofactor. In all hitherto studied class I RNRs, the catalytic radical is generated and stabilized by a dinuclear metal site in protein R2 in an oxygen-dependent reaction, and then reversibly shuttled to protein R1 where ribonucleotide reduction occurs^{8–10}. The dinuclear metal site is coordinated by four carboxylate residues and two histidines. Depending on subclass, the cofactor is: di-iron (class Ia), di-manganese (class Ib) or heterodinuclear Mn/Fe (class Ic)^{11,12}. Classes Ia and Ib generate a stable tyrosyl radical, whereas proteins of class Ic form a radical-equivalent, Mn(IV)/Fe(III) high-valent oxidation state of the metal site^{13,14}. Class Id, containing a Mn(IV)/Mn(III) cofactor, has also been proposed recently^{15–17}. The metal sites in classes Ia and Ic perform direct oxygen activation, whereas class Ib requires a flavoprotein, NrdI, to generate superoxide that oxidizes the di-manganese site^{18–20}.

Sequence analysis revealed a group of class I RNR operons that are present in common human pathogens, for example *Mycoplasma genitalium*, *Mycoplasma pneumoniae* and *Streptococcus pyogenes*. Analogous to standard class Ib RNRs, the operons contain the genes *nrdE*, *nrdF* and *nrdI*, which encode the proteins R1, R2 and NrdI, respectively.

Phylogenetically, the group forms a clade derived from class Ib proteins (Extended Data Fig. 1). Notably, the R2 proteins in this group retain only three of the six metal binding residues; in all other known R2 subclasses, all six residues are completely conserved and are each essential (Fig. 1a). These substitutions seem to exclude a metal site and a radical generation mechanism that is even remotely similar to any ribonucleotide reductase studied so far. In many cases the valine, proline and lysine (VPK) or glutamine, serine and lysine (QSK) variants represent the only aerotolerant RNR found in the genome, for example in *M. genitalium* and *M. pneumoniae* (VPK) and *Gardnerella vaginalis* (QSK) (Fig. 1b).

We investigated whether a VPK variant operon could rescue an *Escherichia coli* strain that lacks aerobic RNR ($\Delta nrdAB\Delta nrdEF$)²¹ and is otherwise unable to grow in the presence of oxygen. A tunable arabinose-induced pBAD plasmid containing the *nrdFIE* operon from *Mesoplasma florum* was constructed and transformed into the $\Delta nrdAB\Delta nrdEF$ strain. Cultures grown under anaerobic conditions were subsequently exposed to oxygen. The *MfnrdFIE* plasmid rescued the knockout strain and enabled growth under aerobic conditions (Fig. 1c). This result is consistent with our previous observations of the *S. pyogenes* *nrdFIE* operon²².

We proceeded to quantify the in vitro activity of the enzyme. We were unable to obtain in vitro RNR activity using the *MfR2* protein expressed separately in *E. coli*. However, purification of *MfR2* after co-expression of the entire *MfnrdFIE* operon under aerobic conditions resulted in a deep-blue-coloured protein that exhibited RNR activity together with *MfR1* (Fig. 1d). This colour and activity were also observed when *MfR2* was co-expressed with only *MfnrdI* under aerobic conditions, whereas co-expression under anaerobic conditions produced an inactive and colourless *MfR2*. Under our assay conditions, the turnover number is 0.18 s^{-1} or >300 for the duration of the assay. Once the *MfR2* protein is activated, *MfNrdI* is thus not required for multiple-turnover activity in vitro. The specific activity was determined to be $275 \pm 7\text{ nmol min}^{-1}\text{ mg}^{-1}$, in line with that of typical class I RNRs^{16,20}.

Incubation of the active *MfR2* protein with hydroxyurea, a radical-quenching RNR inhibitor, yielded a colourless and inactive protein. The activity of the quenched protein could be partially restored if *MfNrdI* was added to the inactivated *MfR2* and the protein mixture was subjected to reduction–oxidation cycles using dithionite and an oxygen-containing buffer (Fig. 1d).

MfR2 and *MfR1* therefore constitute an active RNR system, but only after *MfR2* has undergone an NrdI- and oxygen-dependent activation step. This is principally similar to class Ib RNRs. Moreover, small angle X-ray scattering (SAXS) measurements showed that *MfR2* and *MfNrdI* form a well-defined 2:2 complex with the same interaction geometry as that of standard class Ib RNR proteins^{19,23} (Extended Data Fig. 2). Notably, however, for class Ib RNR, the role of NrdI is to provide an oxidant for the di-manganese metal site that subsequently generates the catalytic tyrosyl radical; this mechanism seems to be implausible for

¹Department of Biochemistry and Biophysics, Stockholm University, Arrhenius Laboratories for Natural Sciences, Stockholm, Sweden. ²Max Planck Institute for Chemical Energy Conversion, Mülheim an der Ruhr, Mülheim an der Ruhr, Germany. ³Cancer Proteomics Mass Spectrometry, Department of Oncology-Pathology, Science for Life Laboratory, Karolinska Institutet, Solna, Sweden. ⁴Research School of Chemistry, Australian National University, Canberra, Australian Capital Territory, Australia. ⁵Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. ⁶These authors contributed equally: Vivek Srinivas, Hugo Lebrette. *e-mail: hogbom@dbb.su.se

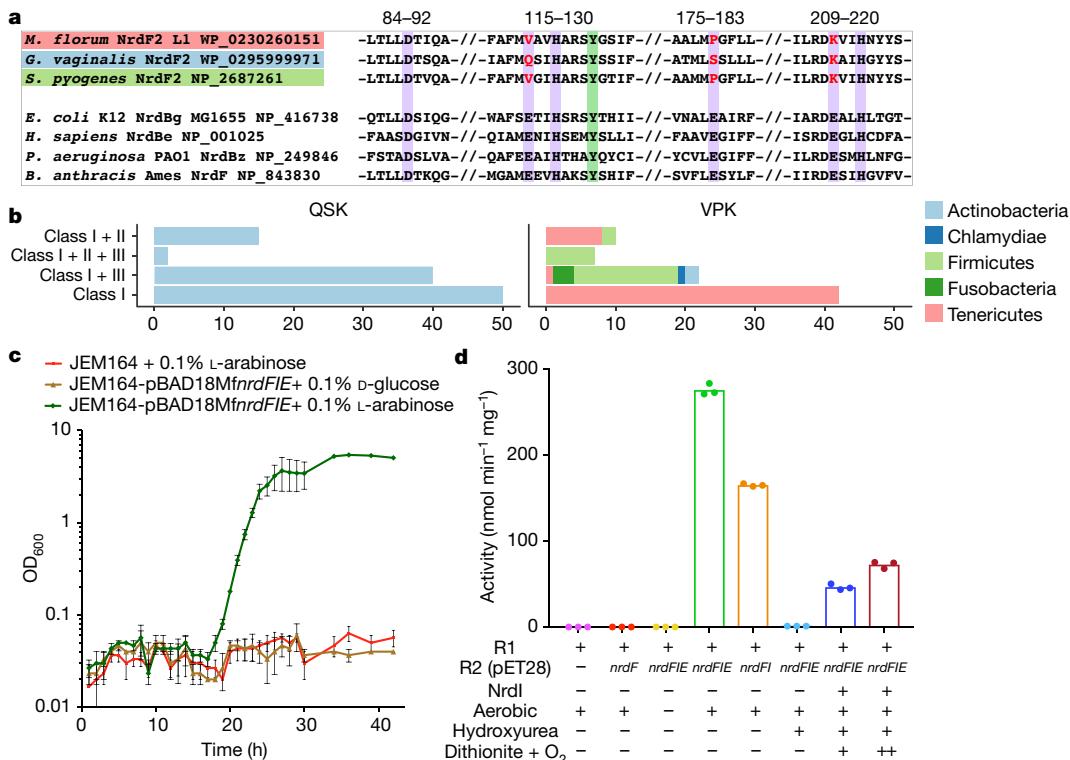


Fig. 1 | A new RNR subclass can rescue an *E. coli* strain that lacks aerobic RNR. **a**, Sequence alignment of the new R2 protein groups to a number of standard, di-metal-containing R2 proteins. Purple background indicates the six metal-binding residues that are normally essential, of which only three are conserved in the new subclass. Two variants are observed in which three carboxylate metal ligands are substituted either for valine, proline and lysine (VPK variant) or for glutamine, serine and lysine (QSK variant). The normally radical-harbouring tyrosine residue is shown with a green background. **b**, The taxonomic distribution of NrdF2, showing QSK-encoding (left) and VPK-encoding (right) organisms and their collected RNR class repertoire. As is common for class I RNRs, several genomes that encode the QSK or VPK variant also harbour other RNRs. The QSK clusters are found only in Actinobacteria, whereas the VPK clusters are also found in Firmicutes, Tenericutes, Chlamydiae and Fusobacteria. **c**, Expression of the *MfnrdFIE* operon induced by

addition of 0.1% v/v L-arabinose (green) rescued the JEM164 double-knockout (Δ *nrdAB* Δ *nrdEF*) strain, whereas when gene expression was suppressed with 0.1% v/v D-glucose (brown) the strain failed to recover, as did the strain lacking the vector (red). Growth curves are shown, data are mean \pm s.d. of three experiments. **d**, *MfnrdI* activates *MfR2* in an oxygen-dependent reaction. High-performance liquid chromatography (HPLC)-based in vitro RNR activity assays show no activity of the R2 protein expressed separately in *E. coli* (red), whereas aerobic co-expression of *MfnrdF* with *MfnrdI* and *MfnrdE* (green) or *MfnrdI* (orange) produced an active R2 protein. Anaerobic co-expression (yellow) or incubation of the active R2 with hydroxyurea (light blue) abolishes the activity. Partial activity could be restored by the addition of *MfnrdI* and redox cycling with dithionite and oxygen for one (blue) and two (maroon) reduction-oxidation cycles. Data points are shown for triplicate experiments.

the *MfR2* and *MfR1* system, given the substitution of the metal-binding residues in *MfR2*.

The crystal structure of *MfR2* in its active form was determined to 1.5 Å resolution, and two crystal structures of the inactive form—either expressed alone aerobically or co-expressed with the entire operon under anaerobic conditions—were resolved to 1.2 Å (Extended Data Table 1). *MfR2* shows the same fold and dimeric arrangement as standard metal-binding R2 proteins (Fig. 2a). As expected from the sequence, the site normally occupied by the dinuclear metal cofactor is markedly atypical (Fig. 2b, c). No electron density corresponding to a metal ion could be observed in this site in any of the structures. We also collected X-ray anomalous scattering data at a wavelength of 0.97 Å. Anomalous difference maps would reveal even low-occupancy metal binding; however, no signal above noise could be observed in the vicinity of the site. To rule out loss of the metal during crystallization, we conducted total-reflection X-ray fluorescence (TXRF) analysis on the active *MfR2* protein solution, to quantitatively detect all elements from aluminium to uranium (with the exception of zirconium, niobium, molybdenum, technetium, ruthenium and rhodium). Only trace amounts of metals could be detected in the active protein sample, with, for example, the following molar ratios of metal to protein: manganese 0.04%, iron 0.35%, cobalt 0.00%, nickel 1.70%, copper 0.99% and zinc 0.71% (Extended Data Fig. 3). Cumulatively, the transition-metal

content corresponded to less than 0.04 per protein monomer. TXRF analyses were also performed for the purified *MfR1* and *MfNrdI* proteins in solution, again with only trace amounts of metals detected.

Notably, the tyrosine seems to be modified in the *meta* position in the active *MfR2* protein but is not modified in the inactive protein (Fig. 2c, d). Mass spectrometry analysis confirmed this covalent modification. The full-length active *MfR2* protein is 17 \pm 2 Da heavier than the inactive protein: 39,804 \pm 2 Da compared with 39,787 \pm 1.4 Da, respectively (Extended Data Fig. 4). Proteolytic cleavage and peptide analysis pinpoints the modification to +15.995 \pm 0.003 Da (monoisotopic mass, corresponding to one additional oxygen atom) at Tyr126 (Extended Data Fig. 5). On the basis of the X-ray crystallographic and mass spectrometry data, we conclude that Tyr126 is *meta*-hydroxylated in the active protein.

The ultraviolet-visible light (UV-vis) absorption spectrum of active *MfR2* has a similar intensity and profile to that of the tyrosyl radical in canonical R2 proteins; however, the absorption maximum is considerably blue-shifted, appearing at around 383 nm compared with around 410 nm for the canonical tyrosyl radical. Moderately intense (3,000 M⁻¹ cm⁻¹) transitions in this region are a fingerprint of an aromatic radical. It has previously been demonstrated in simpler phenoxyl-radical model systems that *meta*-OH substitution leads to a blue shift of the radical π - π^* marker band²⁴. Incubation of the protein with hydroxyurea led

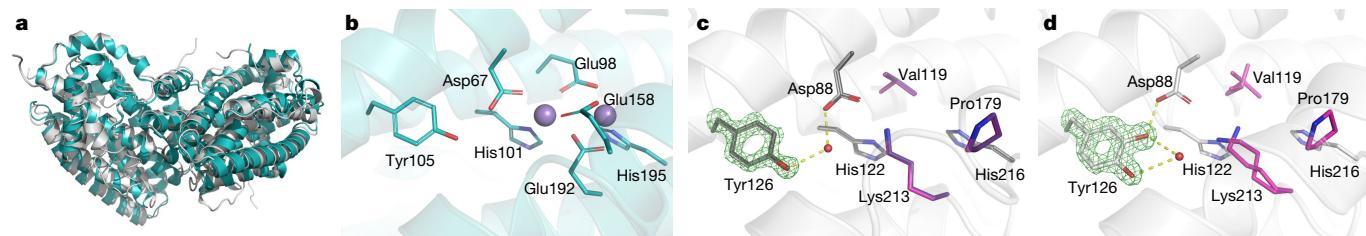


Fig. 2 | The active R2 protein is metal-free but covalently modified. **a**, Overall structure of the *M. florum* VPK R2 protein (grey) compared to the standard class Ib R2 from *E. coli* (PDB ID: 3n37) (cyan). **b**, Structure of the dinuclear metal site and the conserved metal-coordinating residues in standard class I RNR R2. **c**, Structures of inactive *MfR2* after expression without *MfNrdI* (grey) or with *MfNrdI* under anaerobic conditions (dark grey), both determined to 1.2 Å resolution. These structures are identical within experimental error. The three residues that substitute the normally conserved carboxylate ligands of the metal site in canonical class I R2 proteins are shown in pink and purple (expression aerobically without and anaerobically with *MfNrdI*, respectively). In the crystal structure of *MfR2*, the canonical metal positions are occupied by (1) a water molecule in a tetrahedral coordination, involving the conserved His216 with distances of 2.8 ± 0.1 Å—this distance is as expected for a hydrogen-bonded water, but very unlikely for a metal; and (2) the ε -amino group of

to the complete disappearance of the absorbance features and produced a colourless protein (Fig. 3a). In the absence of radical quenchers the features are very stable, with no observable decay after 400 min at 25 °C (Extended Data Fig. 6a). The incorporation of deuterium-labelled amino acids showed that the stable radical resides on a tyrosine-derived residue (Extended Data Fig. 6b).

Electron paramagnetic resonance (EPR) and electron nuclear double resonance (ENDOR) spectroscopy gave results consistent with a DOPA radical, but not an unmodified tyrosyl radical (Fig. 3b, c; for a full description see Extended Data Fig. 7 and Supplementary Information). Spectral simulations suggest that the new substituent is an oxy group (O-X) and are fully consistent with a hydroxyl substituent or a strong hydrogen bond, as indicated by crystallography and mass spectrometry.

Unlike typical R2 proteins, EPR saturation data also demonstrate that the radical has no interaction with a metal (Extended Data

Lys213, replacing the conserved metal-bridging glutamate that is present in all class I R2 proteins. This lysine forms a hydrogen bond with Asp88, the only remaining carboxylate residue. Asp88 also interacts through a hydrogen-bonded water with Tyr126, corresponding to the tyrosine harbouring the metal-coupled radical in standard class Ia and Ib R2 proteins. **d**, Structure of the active *MfR2* after aerobic co-expression with *MfNrdI* and *MfR1*. Here, the tyrosine is covalently modified in the *meta* position; mass spectrometry confirmed that it is hydroxylated. Simulated annealing omit $F_0 - F_c$ electron density maps for the unmodified (**c**) or modified Tyr126 (**d**) are shown in green and contoured at 8σ . In **b**, carbons are in cyan and Mn(II) ions are represented as purple spheres. In **c** and **d**, carbons are shown in grey and pink and hydrogen-bond interactions to Tyr126 are indicated. In **b–d**, oxygens and nitrogens are coloured red and blue, respectively.

Fig. 8), which is consistent with the absence of a metal cofactor^{8,25,26}. Quantification of the radical species by EPR shows that the active *MfR2* sample contains around 52% of radical per protein monomer. As described above, the cumulative amount of transition metals measured by TXRF for the same sample is less than 4% per protein. It is therefore not possible that a metal ion is required to stabilize the observed radical species.

Comprehensive structural, EPR, UV-vis absorption, TXRF and mass spectrometric data thus support the hypothesis that the novel radical species is metal-independent, located on a *meta*-hydroxylated tyrosyl residue and represents an intrinsic DOPA radical within an RNR system.

To investigate whether the observed radical species is catalytically competent, we used the mechanism-based inhibitor 2'-azido-2'-deoxycytidine-5'-diphosphate (N₃-CDP). Incubation of RNR with

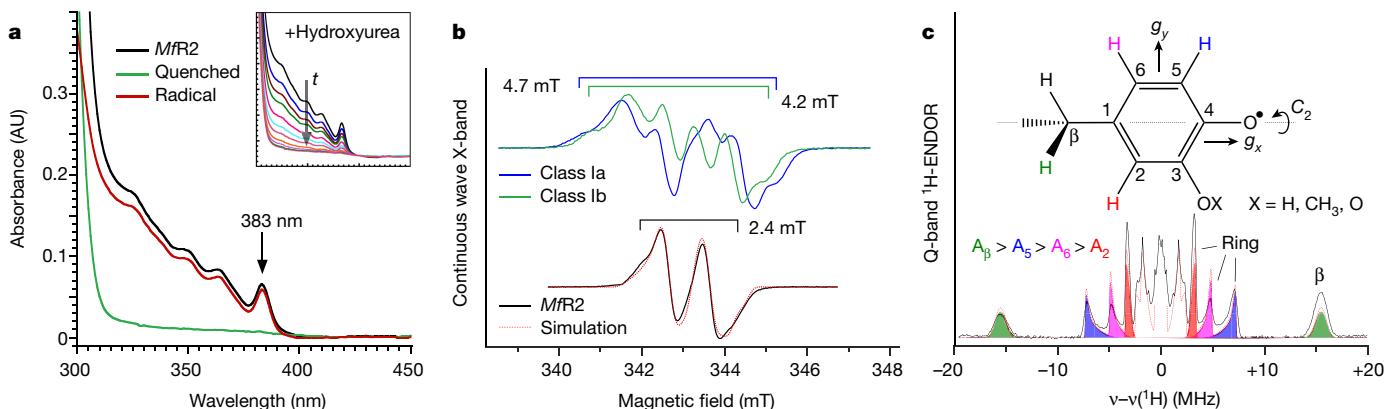


Fig. 3 | Characterization of a stable DOPA radical species in *MfR2*. **a**, The UV-vis spectrum of the active blue-coloured protein shows a peak at 383 nm and additional structure at lower wavelengths (black trace). Incubation with 52 mM hydroxyurea for 20 min removed all features from the spectrum except the protein-related absorbance peak at 280 nm (green trace and inset). The red trace represents the spectrum of the active protein minus the spectrum of the quenched protein. The rate constants for the decay of the absorbance at 348, 364 and 383 nm were identical, which is consistent with all absorbance features arising from a single radical species. **b**, X-band EPR spectra of tyrosyl radicals observed in R2 proteins (*E. coli* class Ia R2 and *Bacillus cereus* class Ib R2 reconstituted with Fe) compared to the *M. florum* R2 radical species reported here. **c**, Q-band ENDOR spectrum recorded at the low-field edge of the EPR

spectrum. The red dashed lines represent a simultaneous simulation of all datasets. Spectral simulations of the multifrequency EPR and ENDOR spectra using the spin Hamiltonian formalism reveal that the radical has four resolved non-equivalent proton couplings with isotropic values of 28.7, 9.8, 6.5 and 4.4 MHz. The absence of a nitrogen coupling excludes the side chains of tryptophan and histidine as the site of the stable radical. In addition, the absence of equivalent proton couplings also excludes the side chains of tyrosine and phenylalanine. Thus, no native aromatic protein residue can explain the observed radical species. The magnitude of the coupling constants is smaller than that of tyrosyl (phenoxy) radicals, and instead is in good agreement with that of phenoxyl radicals with an additional oxy substituent (O-X). All UV-vis, EPR and ENDOR experiments were repeated on three independent protein samples.

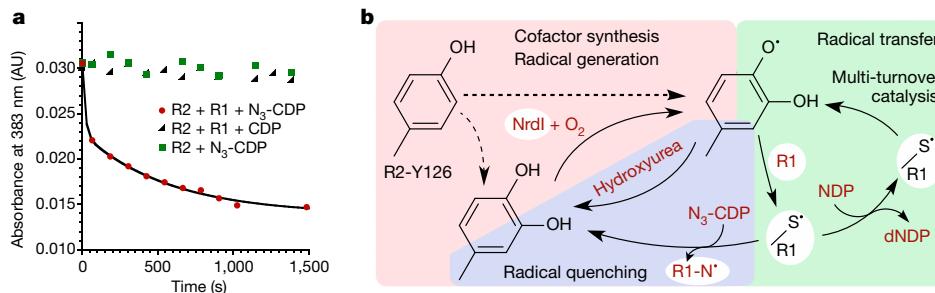


Fig. 4 | Catalytic competency of the radical and proposed mechanistic scheme in class Ie RNR. **a**, UV-vis absorbance at 383 nm plotted against time. The radical signal is quenched in the presence of protein R1 and the mechanism-based inhibitor N₃-CDP (red). Protein R2 with N₃-CDP alone (green) or turnover conditions with protein R1 and CDP (black) does not quench the *Mf*R2 radical. The experiment shows that the observed radical can be reversibly transferred to the active site and support catalysis in protein R1. Experiments were repeated three times. **b**, Proposed mechanistic steps in class Ie RNR. The radical-harbouring cofactor is first

N₃-CDP under turnover conditions leads to loss of the R2 radical, concomitant with the formation of a nitrogen-centred radical in the active site of protein R1^{27–29}. Incubation of active *Mf*R2 with *Mf*R1 and N₃-CDP led to the disappearance of the R2-centred radical species. Catalytic turnover with the substrate (CDP) or incubation of protein R2 with N₃-CDP in the absence of protein R1 did not lead to loss of the R2 radical (Fig. 4a). These results prove that the observed radical in protein *Mf*R2 initiates the catalytic radical chemistry in protein R1 in this new RNR subclass.

Here we identify a type of class I RNR, found in human pathogens, that performs aerobic multiple-turnover ribonucleotide reduction using a DOPA initiator radical. We propose that this RNR group is denoted class Ie. Figure 4b summarizes our current understanding of the system. Notably, NrdI achieves two different oxygen-dependent reactions in this system—tyrosine hydroxylation and DOPA-radical generation—which are two- and one-electron oxidations, respectively. Detailed characterization of the covalent modification and radical generation mechanisms are exciting future prospects.

It has previously been shown that a synthetically introduced DOPA residue in the radical transfer path in *E. coli* RNR functions as a radical trap that prevents RNR turnover; this is presumably because the reduction potential of DOPA is around 260 mV lower than that of a tyrosine at pH 7.0³⁰. The EPR characteristics of the stable radical in *Mf*R2 show that it is electronically more similar to a substituted tyrosyl radical than an *o*-semiquinone. It seems likely that the asymmetry is induced by the protein to tune the redox potential to enable reversible transfer to the cysteinyl radical in the R1 subunit.

Ribonucleotide reductase was the first enzyme found to use a protein-derived radical for catalysis^{26,31}. Although they use different systems for the generation of radicals, all types of RNR in all domains of life were thought to have an absolute metal dependence in order to generate the essential catalytic radical. Differences in metal requirement among RNRs, of which many organisms possess more than one, are believed to provide an advantage in environments in which some metals are limiting. Metal starvation, including the restriction of access to both iron and manganese, is a central strategy in innate immunity to combat invading pathogens³². It is tempting to speculate that class Ie RNR evolved in response to such extremely metal-restricted environments. The post-translational modification of Tyr126 that generates the intrinsic DOPA cofactor is dependent on both NrdI and oxygen, and may potentially also require other factors, including metals. Still, even if this were the case, it would require only catalytic amounts of metal relative to the R2 protein, which would substantially reduce the amount of metal required for the RNR machinery in the organisms that encode this type of RNR.

post-translationally generated by hydroxylation of Tyr126 in an NrdI- and oxygen-dependent process. Dashed lines indicate alternative paths for the post-translational modification of Tyr126. It is currently unknown if this reaction also directly forms the radical species. Once the DOPA radical is formed in protein R2 it supports multiple-turnover ribonucleotide reduction together with protein R1, presumably analogous to other class I RNR systems. If the radical is lost, activity can be restored in the covalently modified R2 protein by NrdI, again in an oxygen-dependent process.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0653-6>.

Received: 7 June 2018; Accepted: 22 August 2018;

Published online 31 October 2018.

1. Hofer, A., Crona, M., Logan, D. T. & Sjöberg, B.-M. DNA building blocks: keeping control of manufacture. *Crit. Rev. Biochem. Mol. Biol.* **47**, 50–63 (2012).
2. Nordlund, P. & Reichard, P. Ribonucleotide reductases. *Annu. Rev. Biochem.* **75**, 681–706 (2006).
3. Mannargudi, M. B. & Deb, S. Clinical pharmacology and clinical trials of ribonucleotide reductase inhibitors: is it a viable cancer therapy? *J. Cancer Res. Clin. Oncol.* **143**, 1499–1529 (2017).
4. Aye, Y., Li, M., Long, M. J. C. & Weiss, R. S. Ribonucleotide reductase and cancer: biological mechanisms and targeted therapies. *Oncogene* **34**, 2011–2021 (2015).
5. Brown, N. C., Eliasson, R., Reichard, P. & Thelander, L. Nonheme iron as a cofactor in ribonucleotide reductase from *E. coli*. *Biochem. Biophys. Res. Commun.* **30**, 522–527 (1968).
6. Lundin, D., Berggren, G., Logan, D. T. & Sjöberg, B.-M. The origin and evolution of ribonucleotide reduction. *Life (Basel)* **5**, 604–636 (2015).
7. Huang, M., Parker, M. J. & Stubbe, J. Choosing the right metal: case studies of class I ribonucleotide reductases. *J. Biol. Chem.* **289**, 28104–28111 (2014).
8. Nordlund, P., Sjöberg, B. M. & Eklund, H. Three-dimensional structure of the free radical protein of ribonucleotide reductase. *Nature* **345**, 593–598 (1990).
9. Uhlin, U. & Eklund, H. Structure of ribonucleotide reductase protein R1. *Nature* **370**, 533–539 (1994).
10. Stubbe, J., Nocera, D. G., Yee, C. S. & Chang, M. C. Y. Radical initiation in the class I ribonucleotide reductase: long-range proton-coupled electron transfer? *Chem. Rev.* **103**, 2167–2202 (2003).
11. Cotruvo, J. A. & Stubbe, J. Class I ribonucleotide reductases: metallocofactor assembly and repair in vitro and in vivo. *Annu. Rev. Biochem.* **80**, 733–767 (2011).
12. Högbom, M. Metal use in ribonucleotide reductase R2, di-iron, di-manganese and heterodinuclear—an intricate bioinorganic workaround to use different metals for the same reaction. *Metalomics* **3**, 110–120 (2011).
13. Högbom, M. et al. The radical site in chlamydial ribonucleotide reductase defines a new R2 subclass. *Science* **305**, 245–248 (2004).
14. Jiang, W. et al. A manganese(IV)/iron(III) cofactor in *Chlamydia trachomatis* ribonucleotide reductase. *Science* **316**, 1188–1191 (2007).
15. Berggren, G., Lundin, D. & Sjöberg, B.-M. in *Encyclopedia of Inorganic and Bioinorganic Chemistry* (ed. Scott, R. A.) <https://doi.org/10.1002/9781119951438.eibc2480> (American Cancer Society, 2017).
16. Rozman Grinberg, I. et al. Novel ATP-cone-driven allosteric regulation of ribonucleotide reductase via the radical-generating subunit. *eLife* **7**, e31529 (2018).
17. Rose, H. et al. Structural basis for superoxide activation of *Flavobacterium johnsoniae* class I ribonucleotide reductase and for radical initiation by its dimanganese cofactor. *Biochemistry* **57**, 2679–2693 (2018).
18. Cotruvo, J. A. Jr, Stich, T. A., Britt, R. D. & Stubbe, J. Mechanism of assembly of the dimanganese-tyrosyl radical cofactor of class I ribonucleotide reductase: enzymatic generation of superoxide is required for tyrosine oxidation via a Mn(II)Mn(IV) intermediate. *J. Am. Chem. Soc.* **135**, 4027–4039 (2013).
19. Boal, A. K., Cotruvo, J. A. Jr, Stubbe, J. & Rosenzweig, A. C. Structural basis for activation of class I ribonucleotide reductase. *Science* **329**, 1526–1530 (2010).

20. Berggren, G., Duraffourg, N., Sahlin, M. & Sjöberg, B.-M. Semiquinone-induced maturation of *Bacillus anthracis* ribonucleotide reductase by a superoxide intermediate. *J. Biol. Chem.* **289**, 31940–31949 (2014).

21. Martin, J. E. & Imlay, J. A. The alternative aerobic ribonucleotide reductase of *Escherichia coli*, NrdEF, is a manganese-dependent enzyme that enables cell replication during periods of iron starvation. *Mol. Microbiol.* **80**, 319–334 (2011).

22. Roca, I., Torrents, E., Sahlin, M., Gibert, I. & Sjöberg, B.-M. NrdI essentiality for class I^b ribonucleotide reduction in *Streptococcus pyogenes*. *J. Bacteriol.* **190**, 4849–4858 (2008).

23. Hammerstad, M., Hersleth, H.-P., Tomter, A. B., Røhr, A. K. & Andersson, K. K. Crystal structure of *Bacillus cereus* class I^b ribonucleotide reductase di-iron NrdF in complex with NrdI. *ACS Chem. Biol.* **9**, 526–537 (2014).

24. Land, E. J. & Porter, G. Primary photochemical processes in aromatic molecules. Part 7.—spectra and kinetics of some phenoxyl derivatives. *Trans. Faraday Soc.* **59**, 2016–2026 (1963).

25. Sahlin, M. et al. Magnetic interaction between the tyrosyl free radical and the antiferromagnetically coupled iron center in ribonucleotide reductase. *Biochemistry* **26**, 5541–5548 (1987).

26. Ehrenberg, A. & Reichard, P. Electron spin resonance of the iron-containing protein B2 from ribonucleotide reductase. *J. Biol. Chem.* **247**, 3485–3488 (1972).

27. Thelander, L., Larsson, B., Hobbs, J. & Eckstein, F. Active site of ribonucleoside diphosphate reductase from *Escherichia coli*. Inactivation of the enzyme by 2'-substituted ribonucleoside diphosphates. *J. Biol. Chem.* **251**, 1398–1405 (1976).

28. Eliasson, R., Pontis, E., Eckstein, F. & Reichard, P. Interactions of 2'-modified azido- and haloanalog of deoxycytidine 5'-triphosphate with the anaerobic ribonucleotide reductase of *Escherichia coli*. *J. Biol. Chem.* **269**, 26116–26120 (1994).

29. Sjöberg, B. M., Gräslund, A. & Eckstein, F. A substrate radical intermediate in the reaction between ribonucleotide reductase from *Escherichia coli* and 2'-azido-2'-deoxynucleoside diphosphates. *J. Biol. Chem.* **258**, 8060–8067 (1983).

30. Seyedsayamdost, M. R. & Stubbe, J. Site-specific replacement of Y356 with 3,4-dihydroxyphenylalanine in the β 2 subunit of *E. coli* ribonucleotide reductase. *J. Am. Chem. Soc.* **128**, 2522–2523 (2006).

31. Sjöberg, B. M., Reichard, P., Gräslund, A. & Ehrenberg, A. The tyrosine free radical in ribonucleotide reductase from *Escherichia coli*. *J. Biol. Chem.* **253**, 6863–6865 (1978).

32. Hood, M. I. & Skaar, E. P. Nutritional immunity: transition metals at the pathogen-host interface. *Nat. Rev. Microbiol.* **10**, 525–537 (2012).

Acknowledgements We would like to acknowledge the foundational contributions of the late P. Reichard to the RNR field. We thank J. Imlay for the gift of the *E. coli* (Δ nrdAB, Δ nrdEF) strain, K. Andersson and M. Hammerstad for the pET-22bBcnrdF plasmid and E. Torrents for the pBAD18 plasmid. Financial support to M.H. was provided by the Swedish Research Council (2017-04018), the European Research Council (HIGH-GEAR 724394), and the Knut and Alice Wallenberg Foundation (Wallenberg Academy Fellows (2012.0233 and 2017.0275)); to B.-M.S. by the Swedish Research Council (2016-01920), the Swedish Cancer Foundation (CAN 2016/670), and the Wenner-Gren Foundations; and to N.C. by the Australian Research Council (FT140100834) and the Max Planck Society. We thank the Diamond Light Source for beamtime (proposals mx11265 and mx15806) and particularly the staff from beamlines I24 and B21.

Reviewer information *Nature* thanks C. Dealwis, M. Fontecave and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions B.-M.S. and M.H. conceived and led the study. D.L. performed bioinformatics. V.S. carried out cloning and generated operon constructs. V.S. and H.L. performed protein production, *in vivo* and *in vitro* activity assays, TXRF analysis, crystallography and sample preparation for all other methods. Y.K., M.S. and N.C. performed spectroscopy. M.L. performed SAXS experiments. J.E. and R.M.M.B. performed mass spectrometry. All authors were involved in experiment design and data analysis. M.H. wrote the manuscript with substantial input from all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0653-6>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0653-6>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

Bioinformatics. The NCBI RefSeq database was searched with custom NrdF HMMER³³ profiles on 17 February 2017 resulting in 4,620 sequences. QSK and VPK variants were manually identified. To reduce the number of highly similar sequences in the phylogeny, the sequences were clustered at 0.75 identity using USEARCH³⁴ and poor-quality sequences were manually removed. The remaining 181 sequences were aligned with ProbCons³⁵ and reliable columns in the alignment were identified with the BMGE program using the BLOSUM30 matrix³⁶. A maximum likelihood phylogeny was estimated from the alignment with RAxML version 8.2.4³⁷, using the PROTGAMMAAUTO model with maximum likelihood estimation starting from a population of 750 rapid bootstrap trees. The number of bootstrap trees was determined with the MRE-based bootstrapping criteria.

Cloning. Genomic DNA of *M. florum* was extracted from a *M. florum* L1 (NCTC 11704) culture obtained from the National Collection of Type Cultures operated by Public Health England. The *M. florum* *nrdF* gene was amplified from *M. florum* genomic DNA by PCR using *M. florum* *nrdF*-forward and *M. florum* *nrdF*-reverse primers and ligated between NheI and BamHI restriction sites of a modified pET-28a plasmid (Novagen), in which the thrombin cleavage site following the N-terminal His6 tag has been replaced by a tobacco etch virus (TEV) protease site. All codons translating to tryptophan in the *MfnrdF* gene were mutated from TGA to TGG to correct for the difference in codon usage between *M. florum* and *E. coli*. The QuikChange Lightning Multi Site-Directed Mutagenesis Kit from Agilent Technologies, using the *M. florum* primers *nrdF*-mut1, *nrdF*-mut2, *nrdF*-mut3 and *nrdF*-mut4 (Extended Data Fig. 9a) as per the protocol suggested by the manufacturers, was used for these changes. Similarly, the pET28MfnrdI plasmid was generated by ligating the *MfnrdI* PCR amplicon resulting from *MfnrdI*-forward and *MfnrdI*-reverse primers into the modified pET-28a plasmid between NdeI and BamHI restriction sites. A synthetic gene coding for *M. florum* *nrdE* with an N-terminal His6 tag and a TEV cleavable site in a pET-21a vector in between the restriction sites NdeI–EcoRI and codon-optimized for overexpression in *E. coli* was ordered from GenScript (pET21MfnrdE). To generate the pET28MfnrdFIE plasmid the *MfnrdF*, *MfnrdI* and *MfnrdE* genes were individually amplified with primer pairs of *MfnrdF*-forward–*MfnrdF*-reverse, *MfnrdI*-forward–*MfnrdI*-reverse and *MfnrdE*-forward–*MfnrdE*-reverse and double-digested by the restriction enzyme pairs NheI–BamHI, BamHI–SacI and SacI–SalI, respectively. The amplicons were then ligated into the NheI and SalI restriction site of a modified pET-28a plasmid with *MfnrdF* placed with an N-terminal TEV-cleavable His-tag and ribosome-binding sites placed ahead of both *MfnrdI* and *MfnrdE* genes (Extended Data Fig. 9b). This entire operon assembly was double-digested from pET28MfnrdFIE and ligated between KpnI and SalI restriction digestion sites of pBAD18 to generate the pBAD18MfnrdFIE plasmid. pET28MfnrdFI was prepared similarly by double-digesting pET28MfnrdFIE with NheI and SacI and ligating it into a modified pET-28a plasmid. All of the abovementioned plasmids were sequenced to check for any unintended mutations.

In vivo studies. A 4-mL LB (ForMedium) culture of *E. coli* double-knockout JEM164 ($\Delta nrdAB\Delta nrdEF\sim srID::Tn10$) supplemented with tetracycline was grown in an anaerobic glove box (MBraun Unilab Plus SP model) with $O_2 < 10$ p.p.m., washed twice with ice-cold water to prepare electrocompetent cells, transformed with pBAD18MfnrdFIE plasmid and plated on a LB-agar medium with tetracycline and carbenicillin. Primary cultures of both JEM164 and JEM164-pBAD18MfnrdFIE were grown anaerobically at 37 °C overnight, supplemented with respective antibiotics. The cultures were removed from the glove box and inoculated in aerobic LB medium supplemented with tetracycline and 0.1% L-arabinose for JEM164 and tetracycline, carbenicillin and either 0.1% D-glucose or 0.1% L-arabinose for JEM164-pBAD18MfnrdFIE and grown for 48 h at 37 °C in a benchtop bioreactor system (Harbinger). Tetracycline was used at a concentration of 12.5 µg mL⁻¹ and carbenicillin was used at a concentration of 50 µg mL⁻¹ for all of the abovementioned cultures.

Protein expression in aerobic conditions. *E. coli* BL21(DE3) (NEB) carrying the plasmid pRAREcamR (Novagen) were transformed with each of the pET28MfnrdF, pET28MfnrdI, pET28MfnrdFI and pET28MfnrdFIE plasmids individually. Glycerol stocks of the transformed cells were flash-frozen and stored at -80 °C. All cultures were grown in LB medium at 37 °C, supplemented with 50 µg mL⁻¹ kanamycin and 25 µg mL⁻¹ chloramphenicol, in a benchtop bioreactor system (Harbinger) until an optical density at 600 nm (OD₆₀₀) of around 0.7 was reached, followed by induction with 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). The cultures were then allowed to grow overnight at room temperature. pET21MfnrdE-BL21(DE3) was grown similarly, but with an antibiotic selection of 50 µg mL⁻¹ carbenicillin instead. Post-expression, the cells were collected by centrifugation and stored at -20 °C until further use.

Protein purification. The cell pellets were thawed and resuspended in lysis buffer (25 mM HEPES-Na pH 7, 20 mM imidazole and 300 mM NaCl) and lysed using a high-pressure homogenizer (EmulsiFlex-C3). The unlysed cells and cell debris was pelleted by centrifugation and the clear supernatant was applied to a lysis-buffer-equilibrated gravity-flow Ni-NTA agarose resin (Protino) column and washed with a lysis buffer with an imidazole concentration of 40 mM. The bound protein was eluted with elution buffer (25 mM HEPES-Na pH 7, 250 mM imidazole and 300 mM NaCl), concentrated and loaded onto a HiLoad 16/60 Superdex 200 size-exclusion column (GE Healthcare) attached to an ÄktaPrime Plus (GE Healthcare) equilibrated with SEC buffer (25 mM HEPES-Na pH 7 and 50 mM NaCl). The fractions corresponding to *MfR2* were pooled and subjected to TEV protease cleavage overnight, at a molar ratio of one TEV protease for every 50 *MfR2* monomers. TEV protease and any uncleaved His-tagged *MfR2* was removed by performing an additional reverse Ni-NTA step. The purity of the protein was evaluated using SDS-PAGE, the protein was then concentrated using Vivapin 20 centrifugal concentrators with a 30,000 Da molecular weight cut-off polyethersulfone membrane (Sartorius) to a desired concentration. All of the above purification steps were performed at 4 °C. The purified *MfR2* protein was then aliquoted, flash-frozen in liquid nitrogen and stored at -80 °C. *MfnrdI* and *MfR1* were purified similarly but with a buffer system of 25 mM Tris-HCl pH 8 instead of 25 mM HEPES-Na pH 7.

Anaerobic *MfR2* production and purification. The glycerol stock of *E. coli* transformed with the pET28MfnrdFIE was used to inoculate a primary culture of LB medium supplemented with 50 µg mL⁻¹ kanamycin and 25 µg mL⁻¹ chloramphenicol and grown in aerobic conditions at 37 °C overnight. All the following steps of protein production and purification were carried out anaerobically in an anaerobic glove box. The primary culture was transferred into the anaerobic chamber and used to inoculate at 2% (v/v) a secondary culture of TB medium (ForMedium) deoxygenated by N₂ saturation and supplemented with 50 µg mL⁻¹ kanamycin and 25 µg mL⁻¹ chloramphenicol, and grown anaerobically at 37 °C until an OD₆₀₀ of 1.1 was reached. This secondary culture was then used to inoculate deoxygenated TB medium supplemented with 50 µg mL⁻¹ kanamycin and 25 µg mL⁻¹ chloramphenicol. As the cell culture reached an OD₆₀₀ of 0.7, overexpression was induced with 0.5 mM IPTG and cells were allowed to grow overnight at 37 °C, the cells were then collected by centrifugation. To extract the soluble fraction, the cell pellet was resuspended at room temperature using 5 mL of deoxygenated BugBuster Protein Extraction Reagent (Novagen) per gram of wet cell paste. The cell suspension was stirred for 40 min at room temperature. The insoluble cell debris was removed by centrifugation at 13,000g for 20 min. From the supernatant, the *MfR2* protein was purified, as described in the aerobic procedure, using deoxygenated buffers. Instead of size-exclusion chromatography, a HiTrap Desalting column (GE Healthcare) was used to remove the imidazole. The protein was then subjected to TEV protease treatment at a molar ratio of one TEV protease for every 10 *MfR2* monomers at room temperature for 1 h. TEV protease and any uncleaved His-tagged *MfR2* was removed by passing the sample over a Ni-NTA agarose (Protino) gravity-flow column equilibrated in the SEC buffer (25 mM HEPES-Na pH 7.0, 50 mM NaCl). The flow-through containing pure cleaved *MfR2* was concentrated to a desired concentration, aliquoted, flash-frozen in liquid nitrogen and stored at -80 °C.

In vitro activity. The in vitro reduction of CDP to deoxy-CDP, catalysed by *M. florum* R1 and R2, was measured in a HPLC system (Agilent 1260 Infinity) using a Waters Symmetry C18 column. A 50 µL reaction mixture consists of 5 µM *MfR1* protein, 5 µM *MfR2* protein, 50 mM of Tris-HCl pH 8.0, 20 mM MgCl₂, 50 mM DTT and 0.5 mM of dATP. The reaction was initiated with the addition of 2 mM CDP and allowed to run at room temperature for 30–60 min, then quenched with 50 µL of methanol. An additional 200 µL of Milli-Q water was added to the reaction mixture. The concentration of the deoxy-CDP product was measured as described previously^{16,38}. All of the in vitro RNR enzymatic reactions were performed with three replicates. Regeneration of quenched *MfR2* in presence of EDTA or metals was performed by incubating *MfR2* (0.3 mM) with either EDTA pH 7.0 (0.3 mM) or a mixture of metals (Mn, Fe, Co, Ni, Cu and Zn at 0.2 mM each), before adding *MfnrdI* and dithionite at 0.6 mM and 2 mM, respectively.

Crystallization. *M. florum* R2 was crystallized using a sitting-drop vapour diffusion method at a protein concentration of 25 mg mL⁻¹. A Mosquito nanolitre pipetting robot (TTP Labtech) was used to set up MRC 2-well crystallization plates (Swissci) with 200:200 nL protein to reservoir volume ratios in a reservoir volume of 50 µL. Initial crystallization hits were obtained in condition number C4 of the PEG/Ion HT crystallization screen (Hampton Research). The crystallization conditions were further optimized using the Additive screen HT (Hampton Research) with condition number B3 producing the best hits. The crystallization conditions are as follows: 100–200 mM calcium acetate, 100 mM ammonium sulfate and 15–20% PEG 3350, with cuboidal crystals appearing within 2–3 days of incubation at room temperature. Reservoir solution supplemented with 20% glycerol was used as a cryoprotectant. Crystals were flash-frozen in liquid nitrogen before data collection.

Data collection and structure determination. X-ray diffraction data were collected at beamline I24 of the Diamond Light Source (Oxfordshire, UK) at a wavelength of 0.96859 Å at 100 K. Data reduction was carried out using XDS³⁹.

MjR2 crystal structures were solved using PHASER⁴⁰ by molecular replacement using the atomic coordinates of the R2 protein from *Corynebacterium ammoniagenes* (PDB ID: 3dhz)⁴¹ as a starting model. A well-contrasted solution was obtained with two molecules per asymmetric unit in the space group C_2 for all the datasets. Crystallographic refinement was performed using PHENIX⁴² applying anisotropic B-factor and TLS, and was similarly but independently conducted for the three models. The 3D models were examined and modified using the program Coot⁴³ and validated using MolProbity⁴⁴. The core root-mean-square deviation values between structures were calculated by the secondary-structure matching (SSM) tool⁴⁵. Extended Data Table 1 was generated with phenix.table_one and lists the crystallographic statistics in which the test set represents 5% of the reflections. The Ramachandran statistics are: favoured 98.5%, 99.2% and 99.5% for the active, inactive aerobic and inactive anaerobic *MjR2* crystal structures, respectively. Ramachandran outliers were 0.0% in all cases. Figure 2 was prepared using the PyMOL Molecular Graphics System, version 2.0 Schrödinger, LLC.

For anomalous signal analysis, data were processed with XDS keeping the Friedel pair reflections unmerged (FRIEDEL'S_LAW = FALSE) to preserve any anomalous contributions. Anomalous difference maps were calculated with PHENIX.

EPR sample preparation. Class Ia *E. coli* *nrdB* was PCR-amplified using primer pairs of *EcnrdB*-forward and *EcnrdB*-reverse from genomic DNA of *E. coli* BL21 (DE3) (Novagen). The PCR amplicon was then double-digested with restriction enzymes pair of NheI-BamHI and ligated into a modified pET-28a plasmid. Class Ib *B. cereus* *nrdF* gene was restriction digested out of pET-22b*BcnrdF* using NdeI-HindIII restriction enzymes and ligated into a modified pET-28a plasmid. The resulting plasmids were sequenced for any unintended mutations. The *E. coli* R2a and *B. cereus* R2b proteins were produced and purified using protocols as described above. A few modifications to the protocols were made in order to obtain metal-free R2 proteins: the addition of 0.5 mM of EDTA to the growth cultures before induction of overexpression with 0.5 mM IPTG, and the addition of 0.5 mM of EDTA to the lysis buffer. The metal-free *E. coli* R2 and *B. cereus* R2 proteins were then metal-loaded with (NH₄)₂Fe(SO₄)₂ solution to a final concentration of 1.5 molar equivalents of Fe(II) per monomer and left to incubate at room temperature for 60 min under aerobic conditions. The R2 proteins were later diluted with 100% glycerol to a final protein concentration of 0.575 mM and 0.695 mM for *E. coli* R2 and *B. cereus* R2, respectively. The active *M. florum* R2 CW X-band EPR sample was also diluted with 100% glycerol to a final protein concentration of 0.675 mM. **Continuous-wave X-band EPR measurements.** X-Band continuous-wave EPR measurements were performed at 70 K using a Bruker E500 spectrometer equipped with a Bruker ER 4116DM TE₁₀₂/TE₀₁₂ resonator, Oxford Instruments ESR 935 cryostat and ITC-503 temperature controller. The microwave power was 20 µW and the magnetic-field modulation amplitude was 0.5 G. The static magnetic field was corrected (for X- and Q-band measurements) using a Bruker ER 035M NMR Gaussmeter. For measurements at 103 K the temperature was controlled by flowing gas of known temperature over the sample in a dewar in the cavity. For spectra at 298 K the sample was contained in a capillary. Quantifications were performed by comparing double integrals of the samples with that a standard Cu(II)/EDTA (1 mM and 10 mM, respectively) sample under non-saturating conditions.

Microwave saturation measurements. The amplitude of the first derivative of an EPR signal is a function of the applied microwave power (P). Plotting the normalized intensity of the EPR signal $X' = (Y'/P^{1/2})/(Y'_0/P_0^{1/2})$ as a function of P in the logarithmic form allows determination of the half-saturation power ($P_{1/2}$)²⁵. Under non-saturating conditions X' equals 1. As the signal starts to saturate, X' decreases. The half-saturation power is calculated from the equation $X' = 1/(1 + P/P_{1/2})^{1/2}$ for an inhomogeneously broadened signal. The value of $P_{1/2}$ is affected by the magnetic environment of the studied radical. The temperature-dependence of $P_{1/2}$ can provide information on an adjacent metal centre, or it may reflect the absence of metals in the vicinity of the radical. R2Tyr⁺ with adjacent di-iron centres have $P_{1/2}$ values that are considerably higher than those for isolated Tyr⁺ radicals at temperatures above 20–30 K²⁵.

EPR radical measurements of *MjR2* with specifically deuterated amino acids. *E. coli* BL21(DE3) (NEB) carrying the plasmid pRAREcamR (Novagen) transformed with the pET28Mf*nrdFIE* plasmid was grown at 37 °C in a benchtop bioreactor system (Harbinger) in M9 minimal medium (Minimal Salts Base from ForMedium containing Na₂HPO₄, KH₂PO₄, NaCl and NH₄Cl) with 4 mM MgSO₄, 0.2 mM CaCl₂, 1% D-glucose and 50 µg ml⁻¹ kanamycin, without the addition of trace metals. The medium was supplemented with either 50 mg l⁻¹ L-tyrosine-(*ring-3,d₂*) or 50 mg l⁻¹ glycine-d₅ or 40 mg l⁻¹ L-tryptophan-(*indole-d₅*) (Cambridge Isotope Laboratories) or 100 mg l⁻¹ D,L-tyrosine-(*β,β-d₂*)⁴⁶. As the cell cultures reached an OD₆₀₀ of 0.6–0.9, 0.5 mM EDTA was added, followed by induction with 0.5 mM IPTG. Induction continued for 36 h at 37 °C. *MjR2* overexpression was checked by

SDS-PAGE. The cells were collected by centrifugation and loaded into EPR tubes before flash freezing. Spectra were recorded at 100 K in a nitrogen flow system at 1 mW and 2 G modulation amplitude. The spectra have been normalized to the same double integrals; that is, the same number of spins in the cavity.

Q-band pulse-EPR measurements. Q-band pulse-EPR measurements were performed in the temperature range of 50 to 70 K using a Bruker ELEXSYS E580 Q-band pulse-EPR spectrometer, equipped with a homebuilt TE₀₁₁ microwave cavity⁴⁷ Oxford-CF935 liquid helium cryostat and an ITC-502S temperature controller. Electron spin echo-detected (ESE) field-swept spectra were measured using the pulse sequence: $t_p - \tau - 2t_p - \tau - \text{echo}$. The length of the $\pi/2$ microwave pulse (t_p) was generally set to 0.5 ns and the interpulse delay τ was 260 ns. Pseudomodulated EPR spectra were generated by convoluting the original absorption spectra with a Bessel function of the first kind. The peak-to-peak amplitude used was 0.8 G. ¹H-ENDOR spectra were collected using the Davis pulse sequence: $t_{\text{inv}} - t_{\text{RF}} - T - t_p - \tau - 2t_p - \tau - \text{echo}$ with an inversion microwave pulse length (t_{inv}) of 140 ns and a radiofrequency pulse length (t_{RF}) of 20 µs (optimized for ¹H). The length of the $\pi/2$ microwave pulse in the detection sequence was set to $t_p = 70$ ns and the interpulse delays to $T = 22$ µs and $\tau = 400$ ns. The radiofrequency was swept 40 MHz around the ¹H Larmor frequency of about 52 MHz (at 1.2 T) in 100 or 50 kHz steps. The radiofrequency amplifier used for ENDOR measurements was ENI 3200L. ¹⁴N-ENDOR spectra were collected using a radiofrequency pulse length of $t_{\text{RF}} = 30$ µs and an interpulse delay of $T = 9.5$ µs. HYSCORE spectra were collected using the pulse sequence: $t_p - t - 2t_p - t_1 - 2t_p - t_2 - t_p - t - \text{echo}$. Two values of t were used, 208 and 220 ns. t_1 and t_2 were incremented in 4-ns steps.

Metal quantification by TXRF. The metal content of protein and buffer solutions was quantified using TXRF analysis on a Bruker PicoFox S2 instrument. For each solution, measurements on three independently prepared samples were carried out. A gallium internal standard at 2 mg l⁻¹ was added to the samples (v/v 1:1) before the measurements. TXRF spectra were analysed using the software provided with the spectrometer. As a control, the metal-free *E. coli* class Ia R2 protein was Fe-reconstituted by incubation with 2 molar equivalents of Fe(II) per monomer at room temperature for 60 min under aerobic conditions. The protein solution was then passed through a HiTrap Desalting column (GE Healthcare) in order to remove unbound iron, and concentrated using a Vivaspin centrifugal concentrator to a concentration similar to that of the *MjR2* sample used for TXRF measurements.

Small angle X-ray scattering. SAXS measurements were carried out on beamline B21 at the Diamond Light Source at 12.4 keV in the momentum transfer range $0.0038 < q < 0.41$ Å⁻¹ ($q = 4\pi \sin(\theta)/\lambda$, 2θ is the scattering angle) using a Pilatus 2M hybrid photon-counting detector (Dectris). Before measurements, samples of *MjR2* alone, or *MjR2* incubated with *MjNrdI* with a molar ratio R2:NrdI of 1:1.5, were run on a HiLoad 16/60 Superdex 200 prep grade size-exclusion chromatography column (GE Healthcare) equilibrated in the SEC buffer 25 mM HEPES pH 7.0, 50 mM NaCl. Fractions corresponding to *MjR2* or to the complex of *MjR2*–*MjNrdI* were pooled, diluted to three different concentrations with the SEC buffer and flash-frozen until measurements. A volume of 30 µl of either *MjR2* alone or *MjR2* incubated with *MjNrdI* was loaded onto the sample capillary using the EMBL Arinax sample handling robot. Each dataset comprised 18 exposures each of 180 s. Identical buffer samples were measured before and after each protein measurement and used for background subtraction. Data averaging and subtracting used the data-processing tools of the EMBL-Hamburg ATSAS package⁴⁸. The radius of gyration (R_g) and maximum particle size (D_{max}) were determined from tools in the PRIMUS program suite⁴⁹. Twenty independent ab initio models of either *MjR2* or the *MjR2*–*MjNrdI* complex were derived from the experimental scattering curves using DAMMIF⁵⁰. For each protein, the models were aligned, averaged and filtered using the DAMAVER program suite⁵¹. Theoretical scattering profiles of the crystal structure of *MjR2* and the *MjR2*–*NrdI* homology model were calculated and fitted against the experimental data using CRYSTOL⁵². The homology model of the *MjR2*–*NrdI* complex was created by superposing the *MjR2* dimer crystal structure and an *MjNrdI* homology model prepared on the basis of the crystal structure of the R2–NrdI complex from *E. coli* (PDB ID: 3N3A)¹⁹.

Proteolytic digestion and peptide analysis by LC-MS. The protein of interest was cleaned from contaminants in the SEC buffer following a modified version of the SP3 protocol^{53,54}. Samples were digested and subjected to LC-MS/MS analysis. Unless noted otherwise, all reagents were purchased from Sigma Aldrich.

The SP3 beads stock suspension was prepared freshly before sample processing. The two SP3 bead bottles (Sera-Mag Speed beads – carboxylate modified particles, P/N 65152105050250 + P/N 45152105050250 from Thermo Scientific) were shaken gently until the suspension looked homogenous, and then 50 µl from each bottle was taken to one tube. The 100 µl of beads was then washed three times with water (in each wash, tubes were taken off the magnetic rack, beads were mixed with 500 µl H₂O by pipetting, then placed back in the rack, and after a 30-s delay, the supernatant was removed). Beads were finally resuspended in 500 µl H₂O after washing.

To each aliquot of 30 μ l of purified protein solution in SEC buffer at 70 μ M protein concentration, 10 μ l of SP3 stock suspension was added. Acetonitrile (MeCN) was added to reach a final concentration of 50% (v/v). The reaction was allowed to stand for 20 min. The beads were collected on the magnetic rack and the supernatant was discarded. The beads were washed twice with 70% (v/v) EtOH and once with MeCN. The beads were then resuspended in 100 μ l of digestion buffer containing either (1) 1 μ g of chymotrypsin in 50 mM HEPES, pH 8, 10 mM CaCl₂, or (2) 1 μ g of pepsin in 50 mM HCl, pH 1.37. Samples were digested at 37°C overnight. The beads were removed and, where necessary, HCl was neutralized with one equivalent of triethylammonium bicarbonate buffer (1 M, pH 8.5). From each sample, 8 μ l was injected to LC-MS/MS analysis on an LTQ-Orbitrap Velos Pro coupled to an Agilent 1200 nano-LC system. Samples were trapped on a Zorbax 300SB-C₁₈ column (0.3 \times 5 mm, 5 μ m particle size) and separated on a NTCC-360/100-5-153 (100 μ m internal diameter, 150 mm long, 5 μ m particle size, picofrit column (Nikkyo Technos) using mobile phases A (3% MeCN, 0.1% formic acid) and B (95% MeCN, 0.1% formic acid) with a gradient ranging from 3% to 40% B in 45 min at a flow rate of 0.4 μ l min⁻¹. The LTQ-Orbitrap Velos was operated in a data-dependent manner, selecting up to three precursors for sequential fragmentation by both collision-induced dissociation and higher-energy collisional dissociation, analysed in the Orbitrap. The survey scan was performed in the Orbitrap at 30,000 resolution (profile mode) from 300–1,700 m/z with a max injection time of 200 ms and AGC set to 1×10^6 ions. MS2 scans were acquired at 15,000 resolution in the Orbitrap. Peptides for dissociation were accumulated for a maximum ion injection time of 500 ms and AGC of 5×10^4 with 35% collision energy. Precursors were isolated with a width of 2 m/z and put on the exclusion list for 30 s after two repetitions. Unassigned charge states were rejected from precursor selection.

The raw files were searched against a database containing only the sequence of *MjR2* protein with Proteome Discoverer 1.4 and the Sequest algorithm⁵⁵ allowing for a precursor mass tolerance of 15 p.p.m. and fragment mass tolerance of 0.02 Da. Non-enzyme searches were conducted and methionine and tyrosine oxidations were included as variable modifications.

For the MODa⁵⁶ searches, raw files were first converted to mzXML files using msconvert from ProteoWizard⁵⁷ using vendor peak picking at the MS² level and otherwise standard settings. The mzXML files were searched against the sequence of *MjR2* with MODa, allowing for an arbitrary number of mods with sizes within the ± 200 Da interval, and using a fragment mass tolerance of 0.05 Da. Otherwise the same settings as for the Sequest searches were chosen. The high-resolution mode was used.

Intact protein analysis by LC-MS. A volume of 2 μ l of inactive *MjR2* and active *MjR2*, at a concentration of 70 μ M, were diluted in 200 μ l of LC-MS mobile phase A. From this, 3 μ l was injected in each LC-MS run, which was set up as described in the previous section, but with the following differences: the trapping column was a Zorbax 300SB-C₈ column (0.3 \times 5mm, 5 μ m particle size); the gradient ramped from 3% B to 99% B in 10 min at a flow rate of 0.6 μ l min⁻¹; the MS only acquired full scans of the 600–4,000 m/z high mass range at 100,000 resolution.

The raw MS files were processed with Protein Deconvolution 4.0 (Thermo Scientific), using the ReSpect algorithm therein. Default parameters were used except: the relative abundance threshold was set to 40%, the m/z range was set to 800–2,400 m/z and the charge state range was set to 6–100.

Data availability

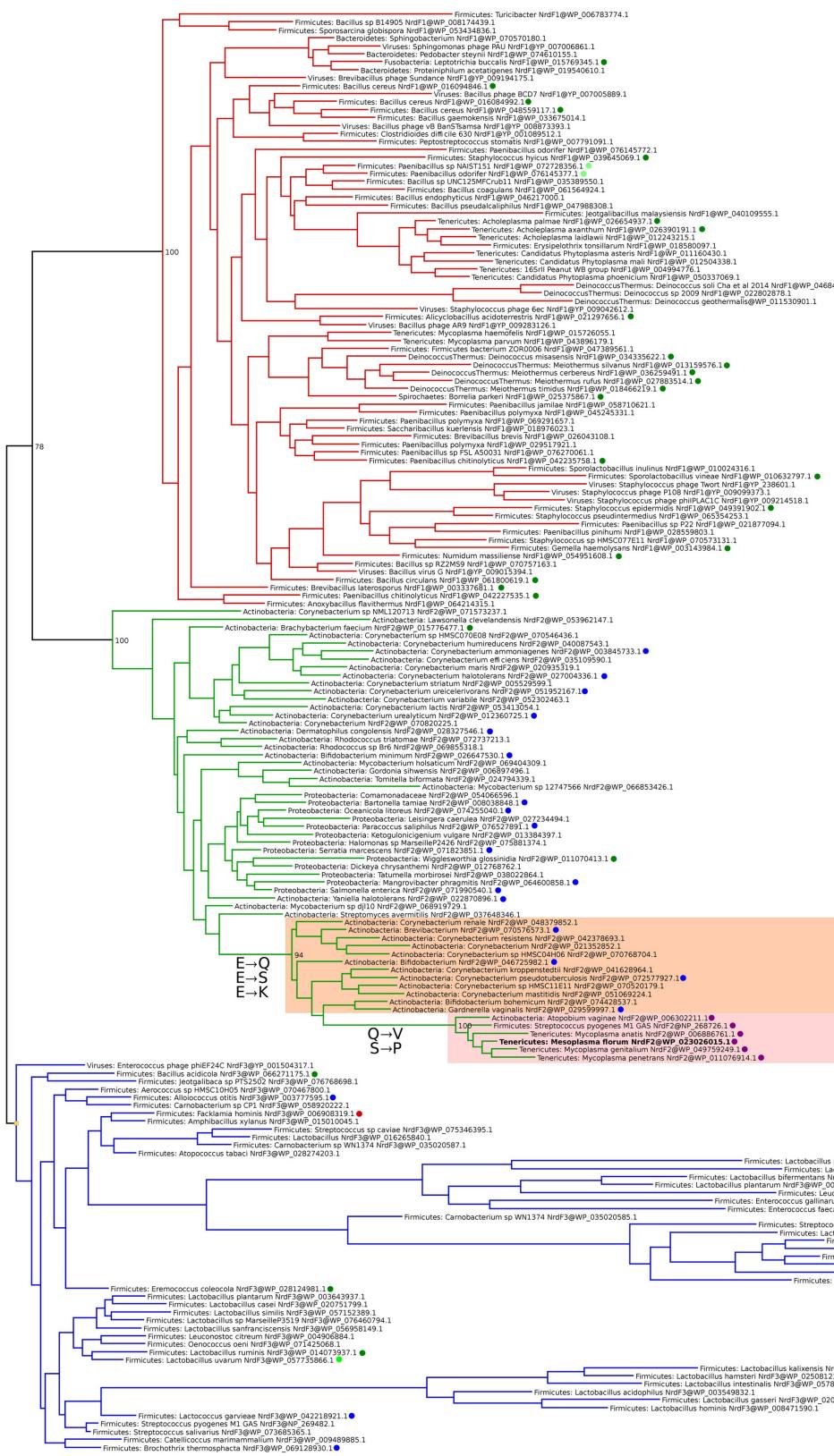
Structures and crystallographic data have been deposited in the Protein Data Bank with the following codes: DOPA-active form, 6GP2; inactive form, 6GP3. All other data can be obtained from the corresponding author upon reasonable request.

33. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
34. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
35. Do, C. B., Mahabhashyam, M. S. P., Brudno, M. & Batzoglou, S. ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.* **15**, 330–340 (2005).
36. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
37. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
38. Loderer, C. et al. A unique cysteine-rich zinc finger domain present in a majority of class II ribonucleotide reductases mediates catalytic turnover. *J. Biol. Chem.* **292**, 19044–19054 (2017).
39. Kabsch, W. XDS. *Acta Crystallogr. D* **66**, 125–132 (2010).
40. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
41. Andersson, M. E. et al. Structural and mutational studies of the carboxylate cluster in iron-free ribonucleotide reductase R2. *Biochemistry* **43**, 7966–7972 (2004).
42. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
43. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
44. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
45. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D* **60**, 2256–2268 (2004).
46. Sjöberg, B. M., Reichard, P., Gräslund, A. & Ehrenberg, A. Nature of the free radical in ribonucleotide reductase from *Escherichia coli*. *J. Biol. Chem.* **252**, 536–541 (1977).
47. Reijerse, E., Lendzian, F., Isaacson, R. & Lubitz, W. A tunable general purpose Q-band resonator for CW and pulse EPR/ENDOR experiments with large sample access and optical excitation. *J. Magn. Reson.* **214**, 237–243 (2012).
48. Franke, D. et al. ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J. Appl. Crystallogr.* **50**, 1212–1225 (2017).
49. Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J. & Svergun, D. I. PRIMUS: A Windows PC-based system for small-angle scattering data analysis. *J. Appl. Crystallogr.* **36**, 1277–1282 (2003).
50. Franke, D. & Svergun, D. I. DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *J. Appl. Crystallogr.* **42**, 342–346 (2009).
51. Volkov, V. V. & Svergun, D. I. Uniqueness of ab initio shape determination in small-angle scattering. *J. Appl. Crystallogr.* **36**, 860–864 (2003).
52. Svergun, D., Barberato, C. & Koch, M. H. CRYSTOL - a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* **28**, 768–773 (1995).
53. Sielaff, M. et al. Evaluation of FASP, SP3, and iST protocols for proteomic sample preparation in the low microgram range. *J. Proteome Res.* **16**, 4060–4072 (2017).
54. Hughes, C. S. et al. Ultrasensitive proteome analysis using paramagnetic bead technology. *Mol. Syst. Biol.* **10**, 757 (2014).
55. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
56. Na, S., Bandeira, N. & Paek, E. Fast multi-blind modification search through tandem mass spectrometry. *Mol. Cell Proteomics* **11**, M111.010199 (2012).
57. Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).

NrdF1

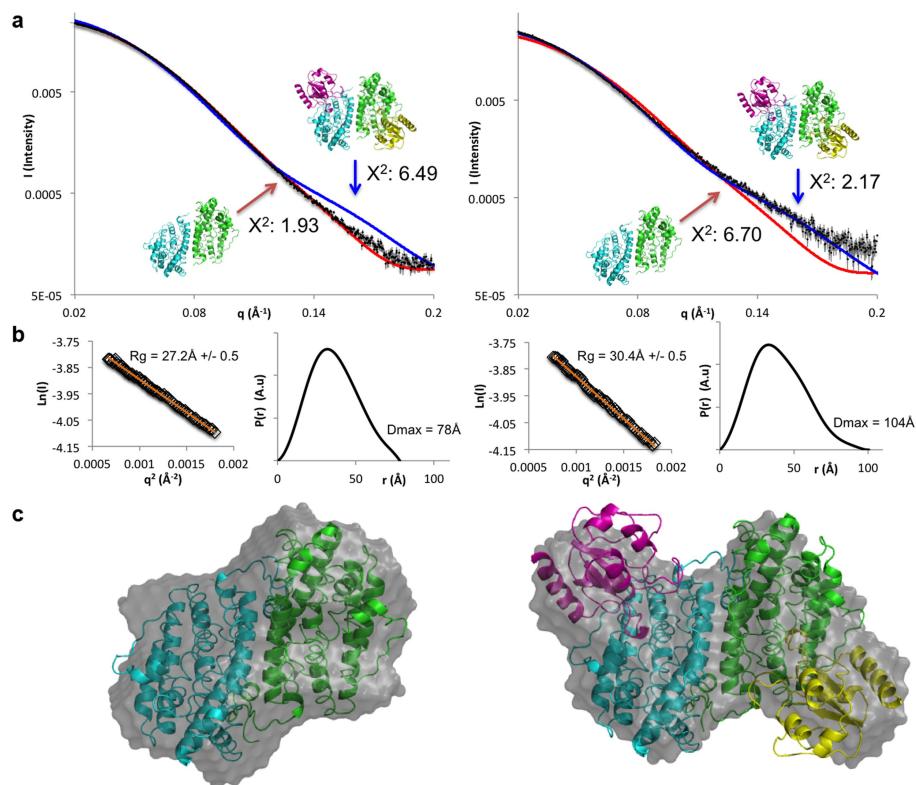
NrdF2

NrdF3



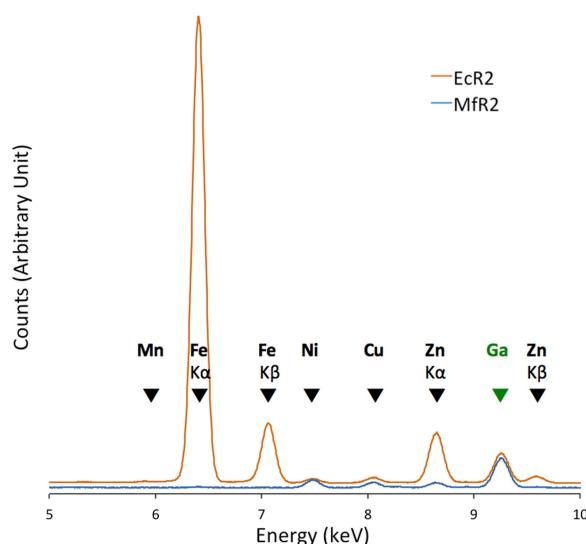
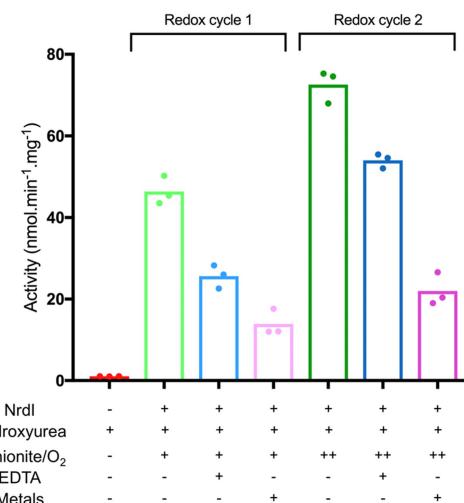
Extended Data Fig. 1 | Unrooted maximum-likelihood phylogeny of representative NrdF (RNR subclass Ib radical-generating subunit) sequences. All RefSeq NrdF sequences were clustered at 75% identity to reduce redundancy and a maximum-likelihood phylogeny was estimated. Sequences with non-canonical amino acids in the positions involved in coordinating the metal centre of the enzyme formed a well-supported clan in the NrdF2 group of sequences. We identified two variants, one in which three of the glutamates were replaced by glutamine, serine and lysine (NrdF2.QSK) and the other in which they were replaced by valine, proline

and lysine (NrdF2.VPK). Both variants thus have a substitution of a lysine for the normally metal-bridging glutamine (residue 213 in *M. florum* NrdF2.VPK). Together, the two variants form a well-supported (96% bootstrap support) clan in the phylogeny inside the NrdF2 diversity. The NrdF2.VPK clan seems to be derived from the NrdF2.QSK clan. Behind the sequences in the tree are a set of sequences that are more than 75% identical to each represented sequence. The VPK and QSK sequences in the phylogeny represent 138 and 182 sequences in RefSeq, respectively.



Extended Data Fig. 2 | Small-angle X-ray scattering characterization of the MfR2-NrdI complex. Solution scattering data for MfR2 (left) and MfR2 incubated with MfNrdI (right). **a**, Experimental solution scattering profiles (black spheres) for MfR2 alone and incubated with MfNrdI superimposed with the theoretical scattering profile of the MfR2 crystal structure (red line) and the theoretical scattering profile from the homology model based on the *E. coli* R2-NrdI complex structure (blue line). Theoretical scattering curves and goodness of fit values were calculated using CRYSTAL. **b**, Guinier fit and $p(r)$ function of MfR2 alone

and incubated with MfNrdI. The fit to the data are shown as an orange line. The shift in invariant parameters R_g and D_{\max} indicate that an increase in dimensions occurred as MfR2 was incubated with MfNrdI. Radius of gyration statistics were derived from 60 data points within the Guinier region for MfR2 and 55 for MfNrdI-MfR2. **c**, Ab initio models (calculated using DAMMIF) of both MfR2 alone and with NrdI (grey surface) overlaid with the crystal structure of MfR2 (left) and the homology model based on the *E. coli* R2-NrdI complex structure model (right).

a**c****b**

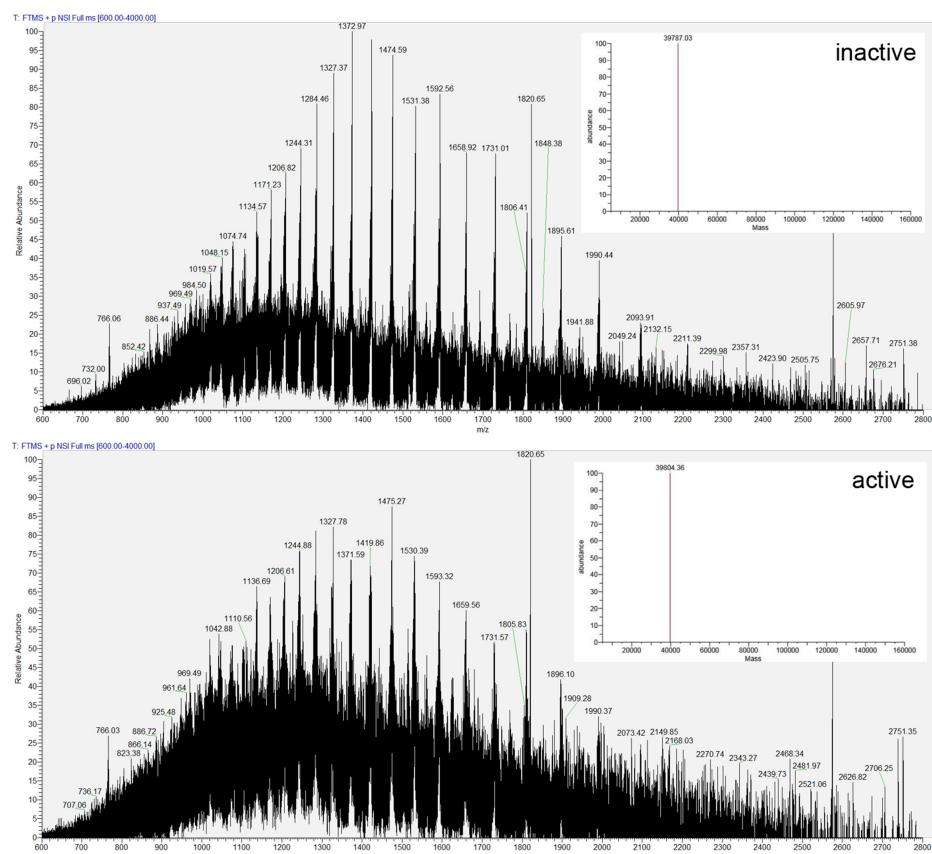
		Mn	Fe	Co	Ni	Cu	Zn	Cumulative
MfR2 active	Conc. (μM)	0,24 ± 0,21	2.29 ± 1.02	n.d.*	11,28 ± 0,27	6,57 ± 0,26	4.71 ± 0,33	-
	mol/mol metal/protein	0.04%	0.35%	-	1.70%	0.99%	0.71%	3.78%
MfNrdI	Conc. (μM)	0.28 ± 0,24	6.84 ± 1.76	n.d.*	3.57 ± 0,38	6.47 ± 0,96	12.20 ± 1.82	-
	mol/mol metal/protein	0.03%	0.76%	-	0.39%	0.71%	1.35%	3.24%
MfR1	Conc. (μM)	1.50 ± 0,21	2.51 ± 0,19	n.d.*	10.50 ± 0,28	4.11 ± 0,04	13.08 ± 0,12	-
	mol/mol metal/protein	0.39%	0.65%	-	2.73%	1.07%	3.41%	8.26%
EcR2a	Conc. (μM)	1.96 ± 0,11	1055.79 ± 40.08	n.d.*	5.44 ± 0,22	6.57 ± 0,09	56.97 ± 2.13	-
	mol/mol metal/protein	0.31%	166.27%	-	0.86%	1.03%	8.97%	177.44%
Buffer 1	Conc. (μM)	0.15 ± 0,15	0.12 ± 0,11	n.d.*	n.d.*	0.04 ± 0,03	0.23 ± 0,02	-
Buffer 2	Conc. (μM)	0.21 ± 0,25	0.19 ± 0,17	n.d.*	0.04 ± 0,07	0.13 ± 0,05	0.20 ± 0,05	-

* n.d. = not detected

Extended Data Fig. 3 | Metal analysis and radical generation in the presence of a chelator. **a**, Representative TXRF spectra measured for *MfR2* (blue, at 664 μM) and Fe-reconstituted class Ia *EcR2* (orange, at 635 μM), on the 5–10 keV energy range. The spectra have been scaled using the peak size of the Ga internal standard and offset slightly in the y direction for clarity. K-level X-ray emission lines are indicated with arrows. For elements, in which both K α and K β lines are present, they are specified. Otherwise, arrows indicate K α lines. Experiments were repeated three times. **b**, Concentrations of Mn, Fe, Co, Ni, Cu and Zn were measured in the active *MfR2*, *MfNrdI* and *MfR1* protein solutions and in their respective buffers, as well as in a solution of *E. coli* class Ia R2 protein reconstituted with Fe *in vitro*. Mean concentrations and s.d. of measurements on three independently prepared samples for each sample are reported. The concentrations were converted to metal-to-protein molar ratio. The measurements show that none of the *MfR* proteins

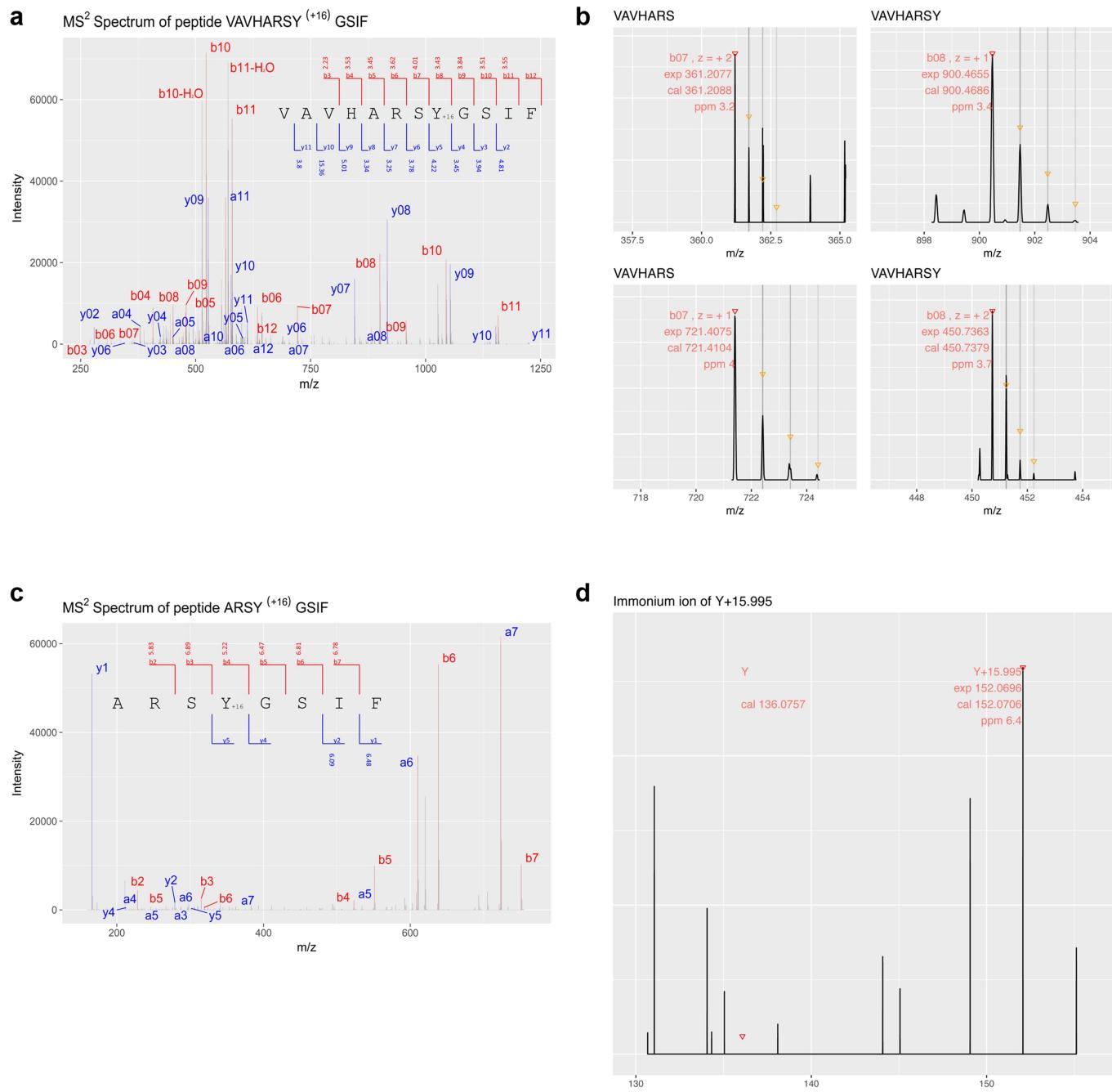
contain a substantial amount of metal, as opposed to *EcR2a* which, as expected, contains in the order of two metal ions per monomer also after a desalting step. Buffer 1 is the buffer system used for *MfR2*; that is, 25 mM HEPES-Na pH 7, 50 mM NaCl. Buffer 2 is the buffer system used for *MfR1* and *MfNrdI*; that is, 25 mM Tris-HCl pH 8, 50 mM NaCl. The protein purification involves a nickel-affinity step, which is probably the reason for nickel being the dominant metal species in the sample.

c, HPLC-based *in vitro* assays show that RNR activity can be restored after *MfR2* is quenched by hydroxyurea. *MfR2* is regenerated by the addition of *MfNrdI* followed by redox cycling with dithionite- and oxygen-containing buffer (green) (see main Fig. 2d). Reactivation and activity are observed also in the presence of a metal chelator (EDTA 0.3 mM, blue). Addition of extra metals (0.2 mM of each Mn, Fe, Co, Ni, Cu, Zn) does not improve the activity recovery (pink).



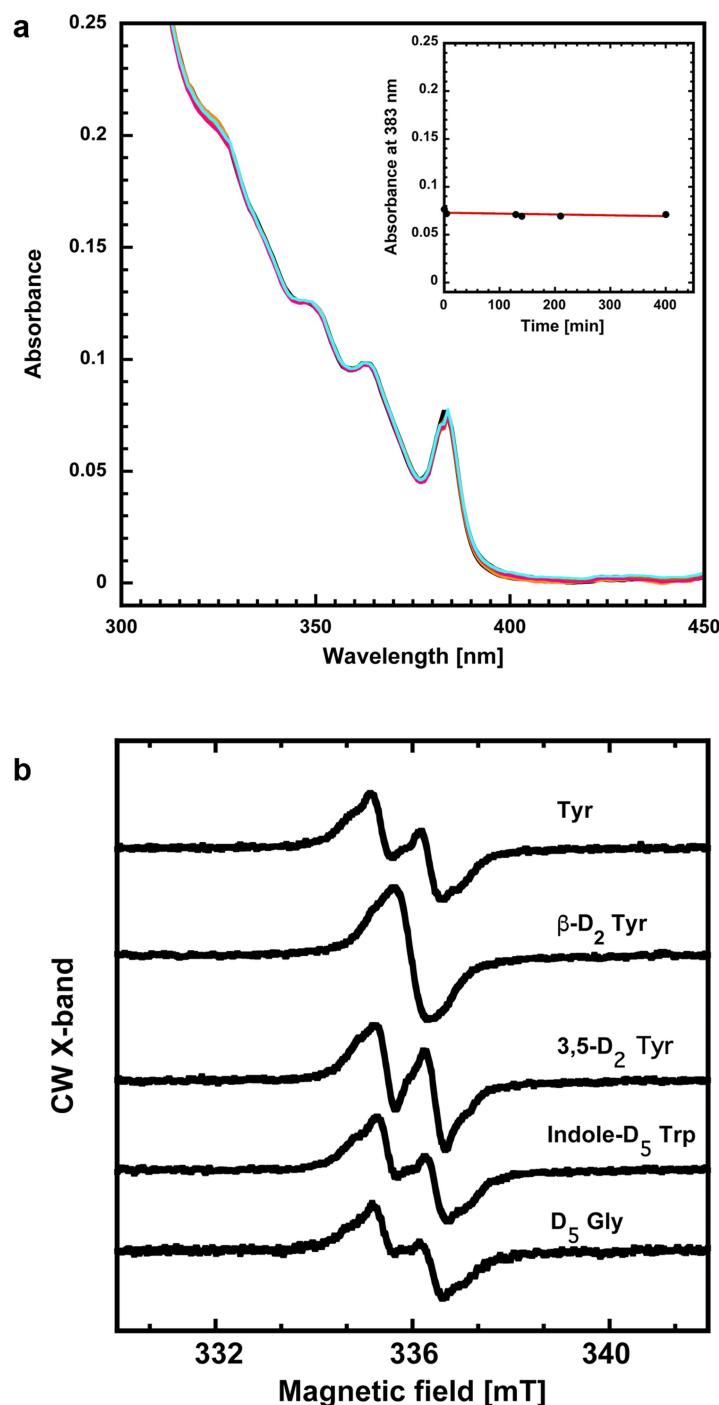
Extended Data Fig. 4 | Mass spectrometric characterization of intact proteins. Intact protein mass spectra obtained from purified *MfR2* proteins. Inactive protein (top) and active protein (bottom). Insets represent the decharged and deisotoped mass as calculated by the program Protein Deconvolution (version 4.0) using the ReSpect algorithm therein. The result from the deconvolution of $n = 30$ consecutive scans in one

LC-MS run per protein form is shown. Each protein form was analysed in duplicate LC-MS runs. Protein intact masses are given as mean \pm s.d. The s.d. was 1.4 Da for the inactive protein and 2 Da for the active form. The results show that the active protein is 17 ± 2 Da heavier than the inactive *MfR2*.



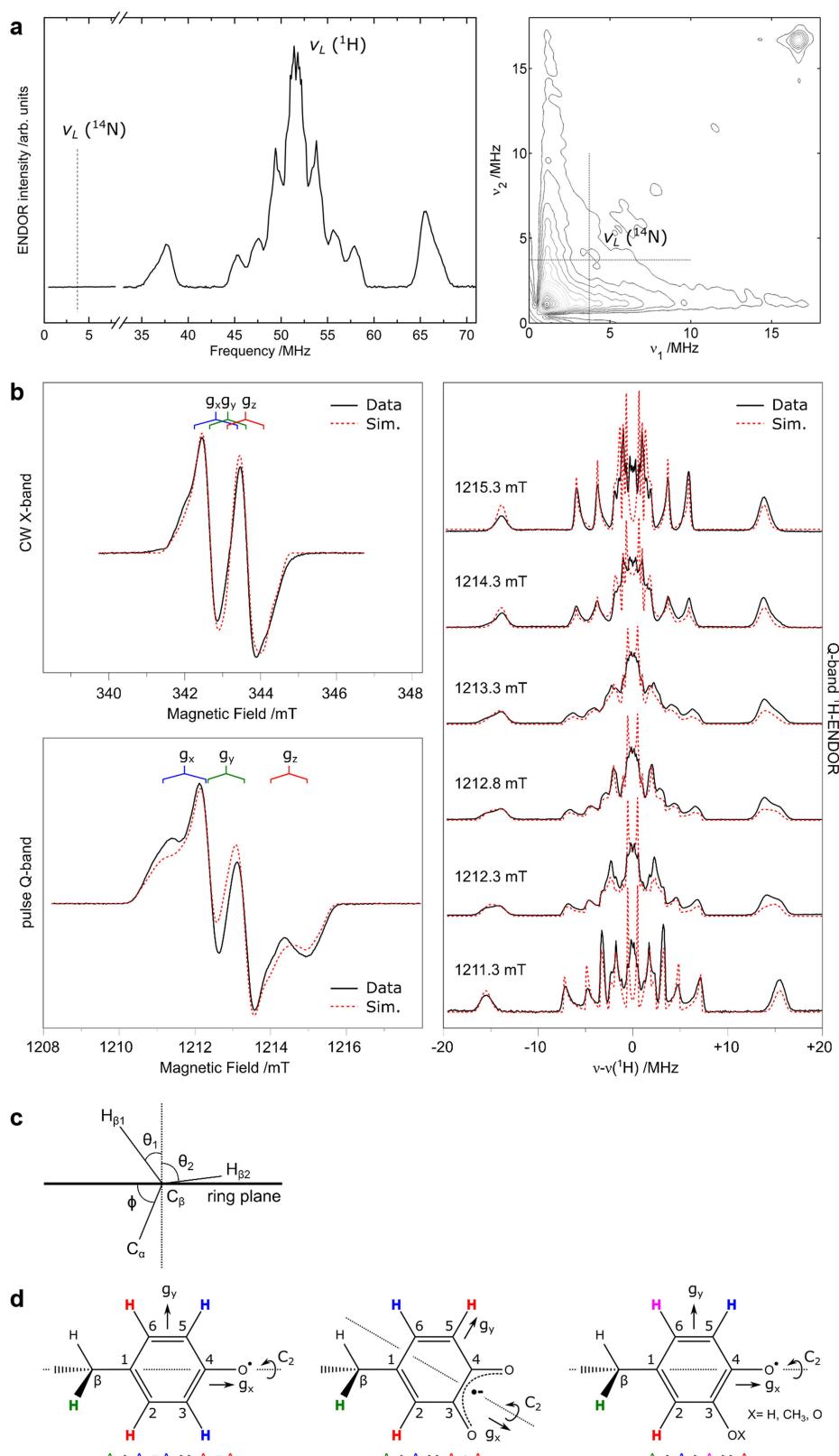
Extended Data Fig. 5 | MS2 fragmentation spectra of peptides with oxidized tyrosine. **a, c**, Annotated MS2 fragmentation spectra and respective theoretical fragment ion tables of the doubly charged precursor ion 661.8458 m/z corresponding to peptide VAVHARSY(+15.995)GSIF (**a**) and the doubly charged precursor ion 458.7279 m/z corresponding to peptide ARSY(+15.995)GSIF (**c**) both with the oxidized (+16) Y126 residue. The peptides shown in **a** and **c** were obtained by proteolytic digestion of the active form of the *MfR2* protein with chymotrypsin and pepsin, respectively. The mass error is typically less than 0.01 m/z , in accordance with the high resolution used (15,000). Errors in p.p.m. are

indicated for the corresponding fragment ions when detected. Among the fragment ions observed, the most relevant are the b7 and b8 ions for the peptide shown in **a**. The experimental m/z values, the annotation, theoretical m/z values and p.p.m. errors are shown in **b**, including the peaks for the corresponding isotope envelope. In **d** the Y(+O) immonium ion for **c** is shown, demonstrating that Tyr126 is modified by a mass of +15.995 and the absence of the corresponding immonium ion for the unmodified Y. Four independent experiments per peptidase treatment were performed, confirming the modified peptide sequences shown. Figures are taken from one representative experiment.



Extended Data Fig. 6 | Radical stability and isotope labelling.
a, Superimposed UV-vis absorption spectra at time points between 0 and 400 min. Inset, absorbance at 383 nm at 0, 4, 129, 140, 210 and 400 min. Experiments were repeated three times. **b**, X-band spectra of the radical observed in collected cells grown in minimal medium supplemented with deuterated amino acids. EDTA (0.5 mM) was added before induction. From top: non-labelled tyrosine, β , β -D₂ tyrosine, 3,5-D₂ tyrosine, indole-D₅ tryptophan and D₅ glycine. The doublet signal collapses to a singlet when β , β -deuterated tyrosine is incorporated in the protein. Furthermore, the

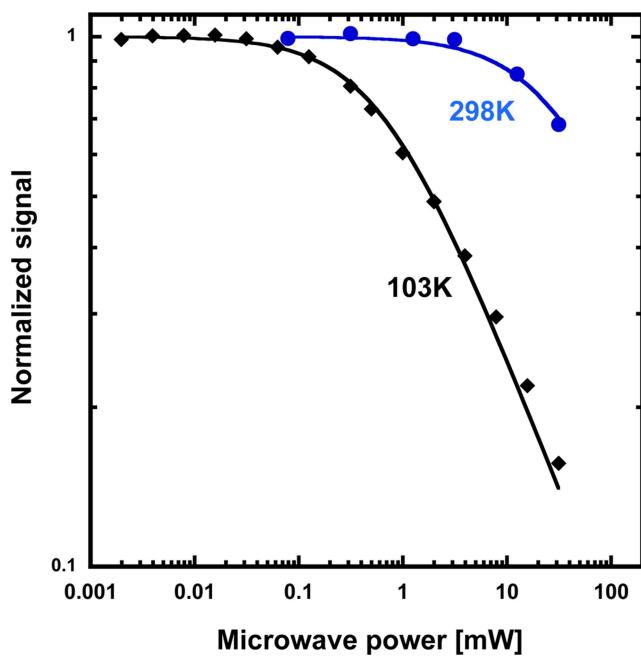
additional coupling to the remaining 3 or 5 proton in the 3,5-d₂ tyrosine grown cells disappears, which is also in line with the radical being tyrosine-derived. Finally, the exclusion of tryptophan and glycine as source for the observed radical is evident from the two lower traces in the figure, which are identical to the top spectrum. Five independent cultures were grown, each including one of the indicated deuterated amino acids. Spectra were recorded at 100 K in a nitrogen-flow system. The spectra have been normalized to the same double integrals; that is, the same number of spins in the cavity.



Extended Data Fig. 7 | EPR and ENDOR characterization.

See Supplementary Information. **a**, Q-band ENDOR and HYSCORE spectra of the spectral region in which an ${}^{14}\text{N}$ hyperfine coupling should be observed. Experimental parameters are listed in Methods. **b**, Full multifrequency (X-band, top left; Q-band, bottom left) EPR dataset and corresponding field dependent Q-band ENDOR spectra (right). Experimental parameters are listed in Methods. The red dashed lines represent a simultaneous simulation of all datasets using the spin Hamiltonian formalism. Simulation parameters are listed in

Supplementary Table 1. **c**, Inferred orientation (θ_1, θ_2) of the C_β protons relative to the phenoxy radical ring plane as determined by the dihedral angle (ϕ) between the ring plane (C_1) and C_α . **d**, Candidates proposed for the $\text{MfR}2$ radical species, see Supplementary Information. All ENDOR measurements were repeated at a second microwave frequency (W-band) giving similar results. Pulse EPR and ENDOR measurements represent extensive data accumulations or averages. EPR: 300 averages (6 scans/50 shots). ENDOR: 600 averages (600 scans/1 shot).

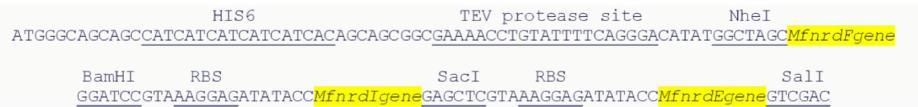


Extended Data Fig. 8 | EPR saturation. EPR saturation behaviour of the *MjR2* radical at 103 and 298 K. Saturation curves at different temperatures determine the microwave power at half saturation $P_{1/2}$. The temperature dependence of $P_{1/2}$ gives information about possible relaxing transition metals in the vicinity of the radical. A fast-relaxing metal site will give a higher $P_{1/2}$ than an isolated radical. The microwave saturation behaviour of the *MjR2* is similar to that for an irradiated tyrosine solution. Here we evaluate $P_{1/2} \approx 0.6$ mW at 103 K and $P_{1/2} \approx 30$ mW at 298 K for *MjR2*. This can be compared to irradiated Tyr^\bullet with $P_{1/2} \approx 0.4$ mW at 93 K and *E. coli* Tyr^\bullet with $P_{1/2} \approx 150$ mW at 106 K and not possible to saturate at 298 K.

a

Primer name	Sequence (5'-3')
MfnrdF-Fow	GTA GCTAGC ATGGCAAAAATAAAAACCAATTACAACGAGTC
MfnrdF-Rev	GTA GGATCC TTAAAACCTCCAATCGTCATCTCAGTTTC
MfnrdF-mut1	GC ^{TT} TCAGTAAACTGG AA CGTAGTAAATGATG
MfnrdF-mut2	TTAGAGGTATGG AA ATAGAATTACACAAA CT TG GG TGCCTG
MfnrdF-mut3	TAAC ^{TT} CATGG AGA ACTTGACACCAGAATGG CA AGAATTAAATTAC
MfnrdF-mut4	GAAGAGGCTCATGAATGG GG TATCAATACAG
MfnrdI-Fow	GTA CATATG CACGATGATATTAAAGTTAG
MfnrdI-Rev	GTA GGATCC TTATTTCCCCAAAATTCTTC ^{AA} ATTTC
MfnrdIO-Fow	GTA GGATCC GTAAAGGAGATATACC ^{AT} GCACGATGATATTAAAGTTAG
MfnrdIO-Rev	GTA GAGCTC TTATTTCCCCAAAATTCTTC ^{AA} ATTTC
MfnrdE-Fow	GTA GAGCTC GTAAAGGAGATATACC ^{AT} GGAGGATAAGAAAATCAATACC
MfnrdE-Rev	GTA GTCGAC TTAGATAACGCACGCATCG
Ecnrdb-Fow	GTA GCTAGC ATGGCATATACCACCTTTACAG
Ecnrdb-Rev	GTA GGATCC TCAGAGTTGGAAGTTACTCAAATCG

Restriction sites are marked in blue and point-mutation sites in red.

b

Extended Data Fig. 9 | Primers and operon construct. **a**, Primers used in this study. **b**, Construction of the *M. florum* class Ie RNR operon.

Extended Data Table 1 | Data collection and refinement statistics

	Active MfR2	Inactive MfR2 (aerobic)	Inactive MfR2 (anaerobic)
Data collection			
Space group	C2	C2	C2
Cell dimensions			
<i>a, b, c</i> (Å)	176.16, 53.74, 79.28	176.19, 53.43, 79.13	176.30, 53.51, 79.10
α, β, γ (°)	90, 108.58, 90	90, 108.53, 90	90, 108.49, 90
Resolution (Å)	42.73-1.48 (1.53-1.48)*	42.71-1.23 (1.27-1.23)	44.19-1.24 (1.28-1.24)
R_{merge}	0.066 (1.403)	0.054 (1.092)	0.052 (0.951)
$I / \sigma I$	12.67 (1.17)	15.13 (1.35)	15.15 (1.50)
Completeness (%)	99.8 (99.7)	98.9 (97.3)	97.4 (94.4)
Redundancy	6.6 (6.4)	6.4 (6.0)	6.5 (6.2)
Refinement			
Resolution (Å)	42.73-1.48 (1.53-1.48)	42.71-1.23 (1.27-1.23)	44.19-1.24 (1.28-1.24)
No. reflections	117311 (11659)	200307 (19622)	192936 (18564)
$R_{\text{work}} / R_{\text{free}}$	0.154 (0.374) / 0.185 (0.408)	0.147 (0.330) / 0.169 (0.343)	0.147 (0.301) / 0.163 (0.311)
No. atoms	6137	6848	6743
Protein	5610	6128	6097
Ligand/ion	2	2	2
Water	525	718	644
B-factors	39.00	27.28	28.61
Protein	38.07	25.71	27.45
Ligand/ion	29.44	20.44	22.58
Water	48.89	40.65	39.67
R.m.s. deviations			
Bond lengths (Å)	0.007	0.006	0.006
Bond angles (°)	1.13	1.12	1.13

A single crystal was used for each dataset.

*Values in parentheses are for highest-resolution shell; the high-resolution cutoff was determined using $\text{CC}_{1/2}$ and refinement behaviour.

Structures of the intermediates of Kok's photosynthetic water oxidation clock

Jan Kern¹, Ruchira Chatterjee^{1,12}, Iris D. Young^{1,12}, Franklin D. Fuller^{1,12}, Louise Lassalle¹, Mohamed Ibrahim², Sheraz Gul¹, Thomas Fransson^{3,11}, Aaron S. Brewster¹, Roberto Alonso-Mori⁴, Rana Hussein², Miao Zhang², Lacey Douthit¹, Casper de Lichtenberg^{5,6}, Mun Hon Cheah⁶, Dmitry Shevela⁵, Julia Wersig², Ina Seuffert², Dimosthenis Sokaras⁷, Ernest Pastor¹, Clemens Weninger⁴, Thomas Kroll⁷, Raymond G. Sierra⁴, Pierre Aller⁸, Agata Butrym⁸, Allen M. Orville⁸, Mengning Liang⁴, Alexander Batyuk⁴, Jason E. Koglin⁴, Sergio Carbojo⁴, Sébastien Boutet⁴, Nigel W. Moriarty¹, James M. Holton^{1,7,9}, Holger Dobbek², Paul D. Adams^{1,10}, Uwe Bergmann³, Nicholas K. Sauter¹, Athina Zouni^{2*}, Johannes Messinger^{5,6*}, Junko Yano^{1*} & Vittal K. Yachandra^{1*}

Inspired by the period-four oscillation in flash-induced oxygen evolution of photosystem II discovered by Joliot in 1969, Kok performed additional experiments and proposed a five-state kinetic model for photosynthetic oxygen evolution, known as Kok's S-state clock or cycle^{1,2}. The model comprises four (meta)stable intermediates (S_0 , S_1 , S_2 and S_3) and one transient S_4 state, which precedes dioxygen formation occurring in a concerted reaction from two water-derived oxygens bound at an oxo-bridged tetra manganese calcium (Mn_4CaO_5) cluster in the oxygen-evolving complex^{3–7}. This reaction is coupled to the two-step reduction and protonation of the mobile plastoquinone Q_B at the acceptor side of PSII. Here, using serial femtosecond X-ray crystallography and simultaneous X-ray emission spectroscopy with multi-flash visible laser excitation at room temperature, we visualize all (meta)stable states of Kok's cycle as high-resolution structures (2.04–2.08 Å). In addition, we report structures of two transient states at 150 and 400 µs, revealing notable structural changes including the binding of one additional 'water', Ox, during the $S_2 \rightarrow S_3$ state transition. Our results suggest that one water ligand to calcium (W3) is directly involved in substrate delivery. The binding of the additional oxygen Ox in the S_3 state between Ca and Mn1 supports O–O bond formation mechanisms involving O5 as one substrate, where Ox is either the other substrate oxygen or is perfectly positioned to refill the O5 position during O_2 release. Thus, our results exclude peroxy-bond formation in the S_3 state, and the nucleophilic attack of W3 onto W2 is unlikely.

All four (meta)stable S-states of photosystem II (PSII) (Fig. 1a, Extended Data Fig. 1a) were populated by illumination of dark-adapted PSII crystals with 0, 1, 2 or 3 flashes (0F–3F; Fig. 1b). The approximately 2 Å resolution (Extended Data Fig. 1b–e, Extended Data Table 1) was sufficient for determining the positions of the oxygens bridging the metal atoms, in addition to the terminal water positions of the Mn_4CaO_5 cluster; the former are critical for discriminating between proposed structures of the S-states, which was not possible in previous structures^{8,9}. Reliable determination of the S-state composition of the PSII crystals obtained by each flash is essential for correct analysis of higher S-state structures. We therefore collected the $Mn K\beta_{1,3}$ emission spectra in situ (see Methods)¹⁰, simultaneously with diffraction data. The first moment of the $K\beta_{1,3}$ peak shifts towards lower energy in response to the first two flashes (0F → 1F → 2F) and to higher energy in the 3F sample (Fig. 1c), as expected from reported Mn redox states^{4,11–14} (Fig. 1a). These data, in combination

with ex situ O_2 evolution measurements, were used to determine the S-state distribution in each illuminated state (see Methods, Extended Data Fig. 2).

Isomorphous difference maps ($F_{\text{obs}} - F_{\text{obs}}$) between the dark and flash-illuminated states at the acceptor side (Fig. 1d–f) indicate a clear

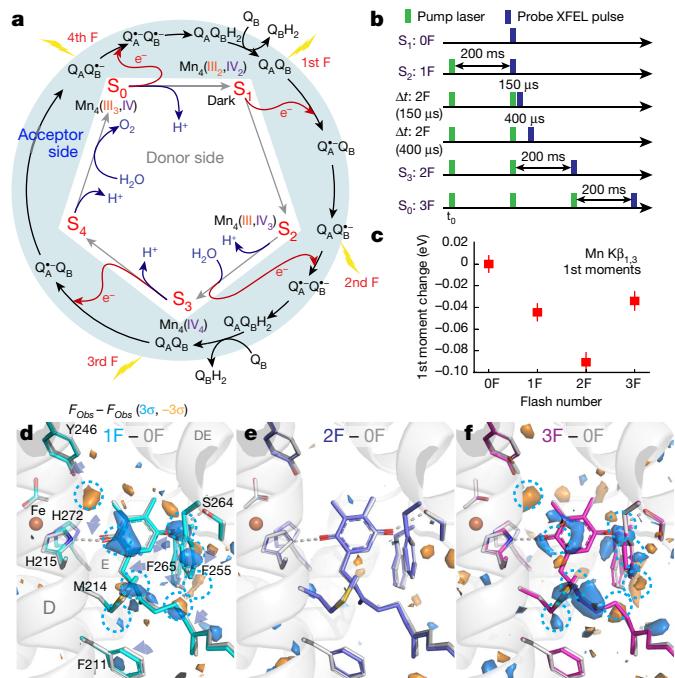


Fig. 1 | The oxygen-evolving cycle in photosystem II. **a**, Relationship between the redox chemistry at the donor (Kok's clock) and acceptor sides throughout the oxygen-evolving cycle. **b**, Relative timing of the visible laser (527 nm) pump and X-ray laser for the different illuminated states. **c**, First moment change of the $Mn K\beta_{1,3}$ XES spectra from PSII crystals obtained in situ simultaneously with XRD. Data shown as mean \pm s.d. (see Methods). **d–f**, $F_{\text{obs}} - F_{\text{obs}}$ isomorphous difference maps around plastoquinone Q_B contoured at $+3\sigma$ (blue) and -3σ (orange) for the 1F, 2F and 3F states relative to 0F. The 0F stick model is shown in light grey; other models have carbons coloured cyan (1F), blue (2F) and magenta (3F).

¹Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ²Institut für Biologie, Humboldt-Universität zu Berlin, Berlin, Germany.

³Stanford PULSE Institute, SLAC National Accelerator Laboratory, Menlo Park, CA, USA. ⁴LCLS, SLAC National Accelerator Laboratory, Menlo Park, CA, USA. ⁵Institutionen für Kemi, Kemiskt Biologiskt Centrum, Umeå Universitet, Umeå, Sweden. ⁶Department of Chemistry—Ångström, Molecular Biomimetics, Uppsala University, Uppsala, Sweden. ⁷SSRL, SLAC National Accelerator Laboratory, Menlo Park, CA, USA. ⁸Diamond Light Source Ltd, Harwell Science and Innovation Campus, Didcot, UK. ⁹Department of Biochemistry and Biophysics, University of California, San Francisco, CA, USA. ¹⁰Department of Bioengineering, University of California Berkeley, Berkeley, CA, USA. ¹¹Present address: Interdisciplinary Center for Scientific Computing, University of Heidelberg, Heidelberg, Germany. ¹²These authors contributed equally: Ruchira Chatterjee, Iris D. Young, Franklin D. Fuller. *e-mail: athina.zouni@hu-berlin.de; johannes.messinger@kemi.uu.se; jyano@lbl.gov; vkyachandra@lbl.gov

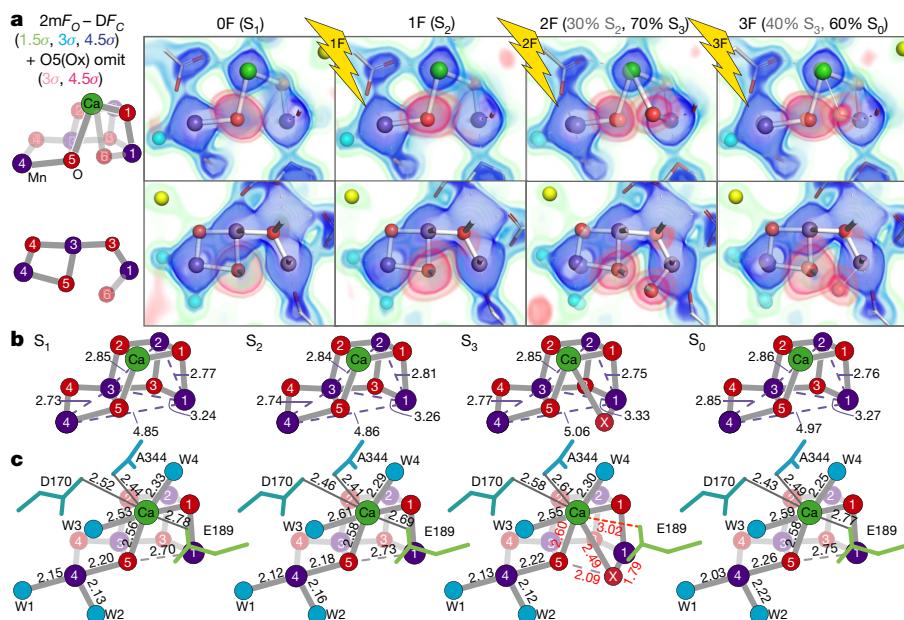


Fig. 2 | Stepwise changes at the OEC during the oxygen-evolving Kok cycle. **a**, Left, labelled diagrams of the OEC atoms in two views. Right, $2mF_{\text{obs}} - DF_{\text{calc}}$ density (green to blue) and O5/Ox omit map density (pink) shown as the overlay of several contour levels for the two views of the OEC in the 0F–3F states. The contributions of the S states to each data set for the two-component analysis are indicated in parentheses.

period-two oscillation of Q_B between its fully oxidized (0F, 2F) and semiquinone $Q_B^{\bullet-}$ (1F) forms. A decrease in the B -factor of the Q_B site after 1F suggests that the quinone in the 1F sample is more tightly bound, owing to the formation of the semiquinone $Q_B^{\bullet-}$. The observation of $Q_B^{\bullet-}$ in the 3F data implies that three electrons were successfully transferred from the Mn_4CaO_5 cluster to the acceptor side, confirming S-state advancement in both PSII monomers (A and a; Fig. 1d–f, Extended Data Fig. 3).

In the S_1 state, the cluster is in a distinct ‘right-open’ structure with no bond between Mn1 and O5; the Mn4–O5 distance is about 2.2 Å, whereas the Mn1–O5 distance is about 2.7 Å (Fig. 2; Extended Data Table 2), with Mn1 clearly pentacoordinate, and likely to be in the +3 oxidation state^{4,11,13–17}. Open, non-cubane structures have been suggested by solution and single-crystal polarized extended X-ray

b, c, Atomic distances in the OEC in each S state in Ångström, averaged across both monomers. Standard deviations for metal–metal, metal–bridging oxygen and metal–ligand distances are 0.1, 0.15 and 0.17 Å, respectively (see Methods). Ca remains 8-coordinate upon insertion of Ox by detaching D1–Glu189.

absorption fine structure (EXAFS) and electron paramagnetic resonance (EPR) data^{4,11,15,18}.

Upon the transition from S_1 to S_2 (1F), one Mn is oxidized from +3 to +4. Figure 2a shows the electron density map ($2mF_{\text{obs}} - DF_{\text{calc}}$) of the dark (S_1) and 1F (predominantly S_2) states. The structure of the cluster in the S_2 state remains fundamentally unchanged, with the coordination numbers of all the metals preserved and in accordance with the similarity of the S_1 and S_2 EXAFS spectra⁴. This ‘right-open’ geometry is consistent with models of the S_2 low spin ($S_{\text{total}} = \frac{1}{2}$) configuration¹⁹, in which Mn4 is +4. There is no indication of a ‘left-open’ structure with an O5–Mn1 bond²⁰. Only small structural changes are observed in the $F_{\text{obs}} - F_{\text{obs}}$ difference map in the oxygen-evolving complex (OEC) region, with shifts of Ca, Mn3 and Mn2 and of residues Asp170 from subunit D1 and Glu354 from subunit CP43 of PS II

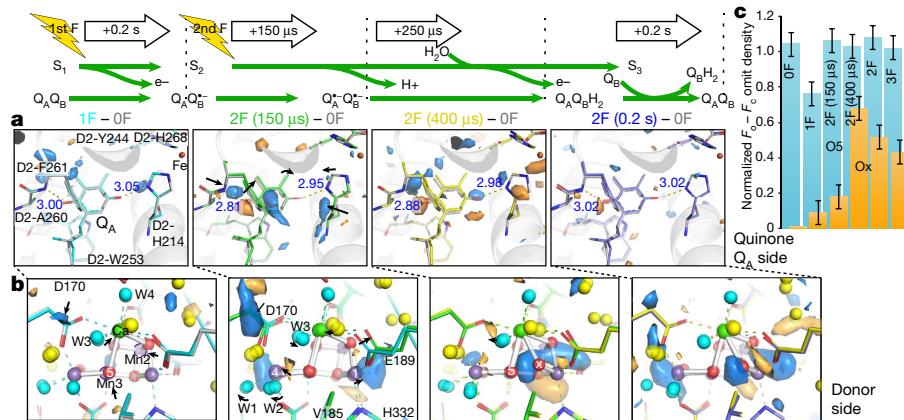


Fig. 3 | The $S_2 \rightarrow S_3$ transition in PSII. **a, b**, Isomorphous difference density at quinone Q_A (**a**) and donor side (**b**) with the expected reduced state of Q_A at 2F (150 μs) and less pronounced at 2F (400 μs), and the oxidation of the OEC and insertion of Ox by 400 μs after the second flash. $F_{\text{obs}} - F_{\text{obs}}$ difference densities between the various illuminated states and the 0F data are contoured at $+3\sigma$ (blue) and -3σ (orange). The model for the 0F data is shown in light grey whereas carbons are coloured

in the models as follows: 1F (cyan), 2F (150 μs) (green), 2F (400 μs) (yellow) and 2F (0.2 s) (blue). **c**, Estimates of occupancies of O5 and Ox based on omit map peak heights normalized against the average electron density maximum at the O2 position in the omit maps. These match full occupancy of O5 throughout and Ox insertion by 400 μs after the second flash. Data shown as mean \pm s.d. based on the electron density value at the O2 position ($n = 12$ observations, see Methods).

(Extended Data Fig. 4). Notably, the strongest difference is observed in the secondary coordination environment: a large negative peak at the location of water W20 (see below).

In Fig. 2a we evaluated the O5 position by overlaying the omit map on the $2mF_{\text{obs}} - DF_{\text{calc}}$ map. In the 0F and 1F data, only one envelope of density is observed. Upon transition from S₂ to S₃ (2F), an additional feature appears near Mn1 that can be assigned to an inserted O atom (hydroxo or oxo). In the $2mF_{\text{obs}} - DF_{\text{calc}}$ map of the 2F data, this additional density is visible as a small but distinct bulge that overlaps with the dominant Mn1 density (Extended Data Fig. 5). The presence of this additional density is distinguishable only in the omit map and in the $2mF_{\text{obs}} - DF_{\text{calc}}$ map at high resolution, and not in the earlier 2.25 Å resolution map⁸ (Extended Data Fig. 5).

We refined the 2F data against two partial-occupancy models at the catalytic site based on the S-state populations (see Methods). The S₂-state structure, which is the refined structure for 1F, was fixed at 30% in the 2F state. The 2F data were then used to refine the S₃-state structure at 70% occupancy. In the refined S₃ structure, the Mn1–Mn4 and Mn1–Mn3 distances are elongated by about 0.2 and 0.07 Å, respectively, relative to the S₂ structure, and the newly inserted hydroxide or oxo (Ox in Fig. 2a, b), located about 1.8 Å from Mn1, occupies its sixth coordination site. This change in coordination of Mn1 from 5- to 6-coordinate is in line with the proposed oxidation of Mn1 from +3 to +4 in the S₂→S₃ transition^{21–23}. Ox is also bound to Ca (2.50 Å) and it is closer to Ca than is O5 (2.60 Å). The Ca coordination number, however, remains eight, as D1-Glu189, which was ligated to Ca in the S₁ and S₂ states (at 2.78 and 2.69 Å, respectively), moves away from Ca in the S₃ state (3.01 Å), making space for Ox (Fig. 2c). These movements are accompanied by changes in the positions of nearby residues (His332, Glu333, His337, Asp342, Ala344, Asp170 of D1, and Glu354, Arg357 of CP43; Extended Data Fig. 4). We note that the current data cannot tell us whether Ox is oxo or hydroxo. However, the position of Ox is compatible with a deprotonated oxo bridge¹⁷.

Suga et al.⁹ recently reported insertion of an O6 atom at 1.5 Å from O5 in their 2F sample (ex situ estimation 46% S₃ state) based on a 2F–0F difference map at 2.35 Å resolution and proposed formation of a peroxide bond between O5 and O6 in the S₃ state. The O5–Ox distance of 2.1 Å in our data is about 0.6 Å longer than the O5–O6 bond modelled by Suga et al.⁹, and the location of Ox is 0.9–1.0 Å away from O6. Thus, our data do not agree with the formation of a peroxide-like bond in the S₃ state. We also note that if a peroxide-like bond were to form in the S₃ state, it would have to be accompanied by reduction of Mn, which conflicts with various spectroscopic observations^{4,15,21} including our current in situ X-ray emission spectroscopy (XES) data (Fig. 1c), which show oxidation of Mn upon the S₂→S₃ transition. We conclude that the differences in interpretation of the data between Suga et al.⁹ and Young et al.⁸ arose from uncertainty when determining oxygen positions at approximately 2.3 Å resolution, whereas our current data clearly show that Ox is bound to Mn1 and Ca, and that there is no peroxo-bond formation with O5 in the S₃ state.

The third laser flash (3F) advances the highest-oxidized metastable S₃ state to the most reduced S₀ state by releasing O₂ and acquiring one water molecule, resetting the catalytic cycle of Kok's clock. The S₀ structure shows the loss of Ox and the return to a motif similar to the dark stable S₁ state (Fig. 2b, c). The decrease in density for Ox in the 3F state (Fig. 2a) is in line with the 60% S₀ population (see Methods).

We collected data sets from two transient states at time points during the S₂→S₃ transition, 150 and 400 μs after the second flash (Fig. 1b), with resolutions of 2.5 and 2.2 Å, respectively, and computed $F_{\text{obs}} - F_{\text{obs}}$ maps with the 0F data (Fig. 3, Extended Data Fig. 6, Supplementary Video 1). At the acceptor side, prominent difference peaks are visible in the vicinity of the primary quinone acceptor, Q_A (Fig. 3a), resembling the differences observed at the Q_B site upon formation of Q_B^{•-} 0.2 s after the first flash (Fig. 1d) and indicating formation of the reduced Q_A^{•-} semiquinone. The same features are visible 250 μs later in the 2F(400 μs)–0F difference map, but at reduced intensity. By contrast, the corresponding $F_{\text{obs}} - F_{\text{obs}}$ difference maps for the data sets collected

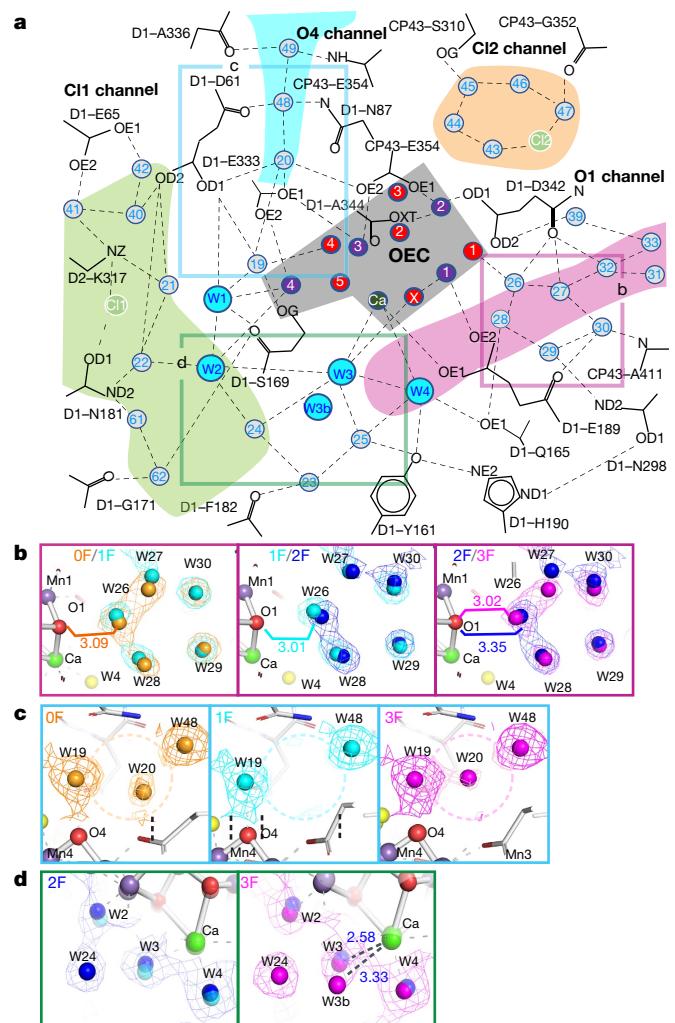


Fig. 4 | Water network around the OEC. **a**, Schematic of the H-bonding network surrounding the OEC indicating starting points of channels connecting the OEC to the solvent-exposed surface of PSII for possible water movement and proton transfer. **b–d**, Changes in selected water positions between the 0F and 3F states overlaid with $2mF_{\text{obs}} - DF_{\text{calc}}$ maps in those states contoured at 1.5σ . Positions in the schematic view are indicated by boxes of the same colour. Selected distances are given in ångström. **b**, Oscillation of the cluster of five waters next to O1 with the O1–O26 distance alternating long–short–long–short from 0F–3F. **c**, W20 and its direct surroundings, indicating disappearance of W20 in the 1F (S₂) data and reappearance in 3F (S₀). **d**, Environment of Ca-bound W3, indicating the presence of a second water position for W3 in the S₀ state (3F).

0.2 s after the first and second flashes do not show any indication of Q_A^{•-} formation, following known kinetics for Q_A^{•-} formation (sub-microsecond) and decay (on the order of hundreds of microseconds)²⁴.

At the OEC, the 2F (150 μs)–0F $F_{\text{obs}} - F_{\text{obs}}$ map shows that the first event after the absorption of the second photon is a movement of Mn4 and Mn1 away from each other by about 0.2 Å (Fig. 3b). Density at the Ox site becomes visible in the 2F (400 μs)–0F $F_{\text{obs}} - F_{\text{obs}}$ map as well as in the $2mF_{\text{obs}} - DF_{\text{calc}}$ map, indicating the ligation of Ox to the Mn1 open coordination site. The Ox density (shown in Fig. 3c with the O5 density) increases substantially at 400 μs, and decreases in the 3F data. The remaining Ox density in the 3F data is explained by the approximately 40% S₃ fraction.

It has been suggested that the S₂→S₃ transition involves a proton transfer followed by an electron transfer^{7,25}. We hypothesize that the proton transfer triggers the shifts of the Mn4 and Mn1 positions in the early stage of the S₂→S₃ transition, including shifts of water positions (see below). One option is that W1 deprotonates, while W3 transfers

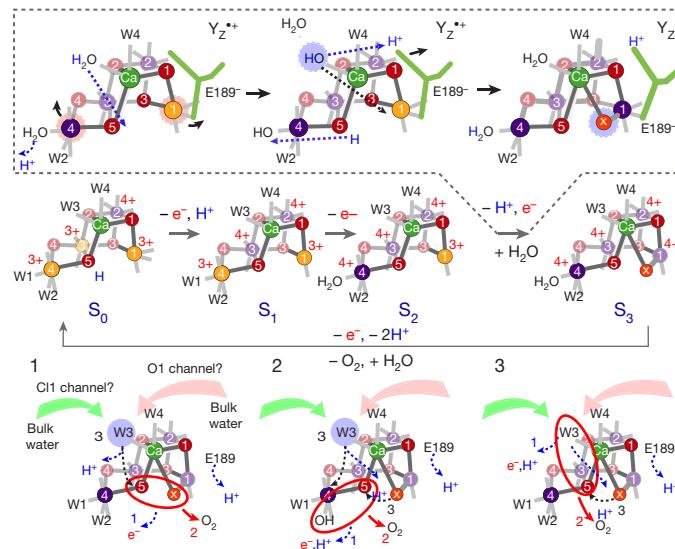


Fig. 5 | Schematic structures of the S states in the Kok cycle of PSII and proposed reaction sequences for O–O bond formation in the S₄ state. The likely position of Mn oxidation states (Mn³⁺ is depicted in orange, Mn⁴⁺ in purple) as well as protonation and deprotonation reactions are indicated for each S state; the proposed steps in the S₂–S₃ transition, including Ox insertion, are indicated in the dashed box with blue dashed arrows signifying atom movements. Three likely options (1, 2 and 3) for the final S₃–S₀ transition are given in the bottom part, including possible order of 1) electron and proton release; 2) O–O bond formation and O₂ release; and 3) refilling of the empty substrate site.

a proton to O5, weakening the O–Mn interaction and allowing the elongation of Mn1 and Mn4²⁶ that is necessary for subsequent oxygen insertion at Mn1, which is coupled to a proton transfer from O5 to W1. D1-Tyr161 (Y_z) is located about 4 Å from D1-Glu189. The formation of the positive charge at Y_z/His189 that precedes oxidation of the OEC could trigger these structural changes in its surrounding, inducing a shift of D1-Glu189 away from Ca as observed in the 2F (150 μs) data. The subsequent oxidation of Mn1 appears to be directly coupled to the insertion of Ox at the Mn1 open-coordination site observed in the 2F (400 μs) data (Fig. 3b, c). We do not observe the formation of a ‘left-open’ structure that has been proposed on the basis of DFT-based studies for the early stage of the S₂–S₃ transition²⁰.

The OEC is embedded in an extended network of H bonds between amino acid residues and waters, which connects it to bulk water and is essential for its function (Fig. 4a, Extended Data Fig. 7, Extended Data Table 3). We observe several substantial changes in this network during the S-state cycle. Movements of the H-bonded water molecules W26–30 (Fig. 4b), of which W26 is located close to O1, indicate that these may be part of a H₂O/H⁺ transfer pathway or take part in charge redistributions within the Mn₄CaO₅ cluster during the S-state cycle. W20 is lost in the S₁–S₂ transition and reappears during the S₃–S₀ transition (Fig. 4c, Extended Data Fig. 7, Supplementary Video 2), suggesting that the O4 channel may be used for proton release in the S₀–S₁ transition but not in the S₂–S₃ and S₃–S₀ transitions. Suga et al.⁹ did not report S₂-state data, but observed the loss of this water in the 2F data and related it to water insertion during S₃-state formation, which was further reinforced by the proposed computational ‘pivot’ or ‘carousel’ mechanisms involving the Mn4 site^{20,27}. However, as this water is already absent in S₂, its loss may not have a direct relationship to the formation of S₃.

Additional water positions are observed near the Ca-bound W3 in the S₀ structure: W3b at 3.25 Å to Ca has approximately 60% occupancy in monomer A (Fig. 4d, Extended Data Fig. 7), while the Ca-bound W3 is at 2.59 Å (approximately 40%). In monomer a, a smaller but similar positive peak is observed in the mF_{obs} – DF_{calc} map. Therefore, we hypothesize that W3 is the entrance site for the water, which is incorporated into the OEC during the S₃–S₀ transition. As indicated

by the changes in the Ca ligation environment, W3 may play a similar role in the S₂–S₃ transition. We therefore suggest that water at the W3 position may act as the entrance site or ‘parking place’ for either the substrate or the next substrate water in the S₂–S₃ and S₃–S₀ transitions, in agreement with earlier suggestions^{5,25,28–30}. Possible access routes of water molecules to W3 are shown in Extended Data Fig. 7 (see also Fig. 4a).

Figure 5 summarizes the structures we determined for all of Kok’s S-states and our interpretation of the S-state-dependent changes with regard to the mechanism of water oxidation. As outlined above, we propose that the Ox site is filled by W3 during the S₂–S₃ transition. The presence of Ox at Mn1, 2.1 Å from O5, suggests that Ox either forms the O–O bond with O5 in the S₃–S₀ transition (case 1 in Fig. 5), or that Ox is placed between Ca and Mn1 to replace O5 during O₂ formation or release. In the latter case, O–O bond formation may occur between O5 and W2 (2) or O5 and W3 (3). We exclude a nucleophilic attack mechanism of W3 or a protein-bound water molecule on W2 or W1, as the basis for these suggestions was that the Mn₃CaO₄ cubane of the Mn₄CaO₅ cluster acts as a ‘battery’ by storing oxidizing equivalents, but remains structurally unmodified in a closed cubane geometry³¹. This is in contrast to our data that show intricate structural changes in an open cubane involving ligand detachment and the formation of a new Ca–Ox–Mn1 bridge.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0681-2>.

Received: 17 April 2018; Accepted: 22 August 2018;

Published online 7 November 2018.

1. Kok, B., Forbush, B. & McGloin, M. Cooperation of charges in photosynthetic O₂ evolution—I. A linear four step mechanism. *Photochem. Photobiol.* **11**, 457–475 (1970).
2. Joliot, P., Barbieri, G. & Chabaud, R. A new model of photochemical centers in system-2. *Photochem. Photobiol.* **10**, 309–329 (1969).
3. Hillier, W. & Wydrzynski, T. ¹⁸O-water exchange in photosystem II: substrate binding and intermediates of the water splitting cycle. *Coord. Chem. Rev.* **252**, 306–317 (2008).
4. Yano, J. & Yachandra, V. Mn₄Ca cluster in photosynthesis: where and how water is oxidized to dioxygen. *Chem. Rev.* **114**, 4175–4205 (2014).
5. Cox, N. & Messinger, J. Reflections on substrate water and dioxygen formation. *Biochim. Biophys. Acta* **1827**, 1020–1030 (2013).
6. Debus, R. J. FTIR studies of metal ligands, networks of hydrogen bonds, and water molecules near the active site Mn₄CaO₅ cluster in photosystem II. *Biochim. Biophys. Acta* **1847**, 19–34 (2015).
7. Klauss, A., Haumann, M. & Dau, H. Seven steps of alternating electron and proton transfer in photosystem II water oxidation traced by time-resolved photothermal beam deflection at improved sensitivity. *J. Phys. Chem. B* **119**, 2677–2689 (2015).
8. Young, I. D. et al. Structure of photosystem II and substrate binding at room temperature. *Nature* **540**, 453–457 (2016).
9. Suga, M. et al. Light-induced structural changes and the site of O=O bond formation in PSII caught by XFEL. *Nature* **543**, 131–135 (2017).
10. Kern, J. et al. Simultaneous femtosecond X-ray spectroscopy and diffraction of photosystem II at room temperature. *Science* **340**, 491–495 (2013).
11. Kulik, L. V., Epel, B., Lubitz, W. & Messinger, J. Electronic structure of the Mn₄Ca cluster in the S₀ and S₂ states of the oxygen-evolving complex of photosystem II based on pulse ⁵⁵Mn-ENDOR and EPR spectroscopy. *J. Am. Chem. Soc.* **129**, 13421–13435 (2007).
12. Dau, H. & Haumann, M. The manganese complex of photosystem II in its reaction cycle—basic framework and possible realization at the atomic level. *Coord. Chem. Rev.* **252**, 273–295 (2008).
13. Peloquin, J. M. et al. ⁵⁵Mn ENDOR of the S₂-state multiline EPR signal of photosystem II: Implications on the structure of the tetrnuclear cluster. *J. Am. Chem. Soc.* **122**, 10926–10942 (2000).
14. Krewald, V. et al. Metal oxidation states in biological water splitting. *Chem. Sci. (Camb.)* **6**, 1676–1695 (2015).
15. Chernev, P. et al. Merging structural information from X-ray crystallography, quantum chemistry, and EXAFS spectra: The oxygen-evolving complex in PSII. *J. Phys. Chem. B* **120**, 10899–10922 (2016).
16. Tanaka, A., Fukushima, Y. & Kamiya, N. Two different structures of the oxygen-evolving complex in the same polypeptide frameworks of photosystem II. *J. Am. Chem. Soc.* **139**, 1718–1721 (2017).
17. Yamaguchi, K. et al. Theory of chemical bonds in metalloenzymes XXI. Possible mechanisms of water oxidation in oxygen evolving complex of photosystem II. *Mol. Phys.* **116**, 717–745 (2018).

18. Yano, J. et al. Where water is oxidized to dioxygen: structure of the photosynthetic Mn₄Ca cluster. *Science* **314**, 821–825 (2006).
19. Pantazis, D. A., Ames, W., Cox, N., Lubitz, W. & Neese, F. Two interconvertible structures that explain the spectroscopic properties of the oxygen-evolving complex of photosystem II in the S₂ state. *Angew. Chem. Int. Ed.* **51**, 9935–9940 (2012).
20. Retegan, M. et al. A five-coordinate Mn(IV) intermediate in biological water oxidation: spectroscopic signature and a pivot mechanism for water binding. *Chem. Sci.* **7**, 72–84 (2016).
21. Dau, H. & Haumann, M. Considerations on the mechanism of photosynthetic water oxidation — dual role of oxo-bridges between Mn ions in (i) redox-potential maintenance and (ii) proton abstraction from substrate water. *Photosynth. Res.* **84**, 325–331 (2005).
22. Cox, N. et al. Photosynthesis. Electronic structure of the oxygen-evolving complex in photosystem II prior to O–O bond formation. *Science* **345**, 804–808 (2014).
23. Siegbahn, P. E. M. Structures and energetics for O₂ formation in photosystem II. *Acc. Chem. Res.* **42**, 1871–1880 (2009).
24. de Wijn, R. & van Gorkom, H. J. Kinetics of electron transfer from Q_A to Q_B in photosystem II. *Biochemistry* **40**, 11912–11922 (2001).
25. Sakamoto, H., Shimizu, T., Nagao, R. & Noguchi, T. Monitoring the reaction process during the S₂→S₃ transition in photosynthetic water oxidation using time-resolved infrared spectroscopy. *J. Am. Chem. Soc.* **139**, 2022–2029 (2017).
26. Rossini, E. & Knapp, E.-W. Protonation equilibria of transition metal complexes: from model systems toward the Mn-complex in photosystem II. *Coord. Chem. Rev.* **345**, 16–30 (2017).
27. Askerka, M., Wang, J., Vinyard, D. J., Brudvig, G. W. & Batista, V. S. S₃ state of the O₂-evolving complex of photosystem II: insights from QM/MM, EXAFS, and femtosecond X-ray diffraction. *Biochemistry* **55**, 981–984 (2016).
28. Boussac, A. & Rutherford, A. W. Nature of the inhibition of the oxygen-evolving enzyme of photosystem II induced by NaCl washing and reversed by the addition of Ca²⁺ or Sr²⁺. *Biochemistry* **27**, 3476–3483 (1988).
29. Tso, J., Sivaraja, M. & Dismukes, G. C. Calcium limits substrate accessibility or reactivity at the manganese cluster in photosynthetic water oxidation. *Biochemistry* **30**, 4734–4739 (1991).
30. Suzuki, H., Sugiura, M. & Noguchi, T. Monitoring water reactions during the S-state cycle of the photosynthetic water-oxidizing center: detection of the DOD bending vibration by means of Fourier transform infrared spectroscopy. *Biochemistry* **47**, 11024–11030 (2008).
31. Barber, J. A mechanism for water splitting and oxygen production in photosynthesis. *Nat. Plants* **3**, 17041 (2017).

Acknowledgements This work was supported by the Director, Office of Science, Office of Basic Energy Sciences (OBES), Division of Chemical Sciences, Geosciences, and Biosciences (CSGB), Department of Energy (DOE)

(J.Y., V.K.Y.), by National Institutes of Health (NIH) grants GM055302 (V.K.Y.), GM110501 (J.Y.) GM126289 (J.K.), GM117126 (N.K.S.), GM124149 and GM124169 (J.M.H.), the Ruth L. Kirschstein National Research Service Award (GM116423-02, F.D.F.), and Human Frontiers Science Project RGP0063/2013 (J.Y., U.B., A.Z.). We acknowledge the DFG-Cluster of Excellence “UniCat” coordinated by T.U. Berlin and Sfb1078, TP A5 (A.Z., H.D.); the Artificial Leaf Project (K&A Wallenberg Foundation 2011.0055) and Vetenskapsrådet (2016-05183) (J.M.); Diamond Light Source, Biotechnology and Biological Sciences Research Council (grant 102593) and a Strategic Award from the Wellcome Trust (A.M.O.). This research used NERSC, supported by DOE, under Contract No. DE-AC02-05CH11231. Synchrotron facilities at the ALS, Berkeley and SSRL, Stanford, were funded by DOE OBES. The SSRL Structural Molecular Biology Program is supported by the DOE OBER, and NIH (P41GM103393). LCLS and SSRL, SLAC National Accelerator Laboratory, are supported by DOE, OBES under Contract No. DE-AC02-76SF00515. We thank the staff at LCLS/SLAC and SSRL (BL 6-2, 7-3) and ALS (BL 5.01, 5.02, 8.2.1, 8.3.1).

Author contributions U.B., V.K.Y. and J.Y. conceived the experiment; R.A.-M., A.Z., J.M., U.B., N.K.S., J.K., V.K.Y. and J.Y. designed the experiment; R.C., M.I., L.L., R.H., M.Z., L.D., J.W., I.S., A.Z. and J.K. prepared samples; A. Batyuk, M.L., S.B., R.A.-M., J.E.K. and S.C. operated the MFX instrument; F.D.F., S.G., E.P., P.A., A.M.O., J.M. and J.K. developed, tested and ran the sample delivery system; M.H.C., D. Shevela, R.C., C.d.L., J.Y. and J.M. performed and analysed O₂ evolution and EPR measurements; R.C., F.D.F., S.G., M.I., C.d.L., M.H.C., I.D.Y., A.S.B., R.A.-M., R.H., M.Z., L.L., D. Sokaras, E.P., C.W., T.F., T.K., R.G.S., P.A., A. Butrym, A. Batyuk, M.L., S.B., J.E.K., S.C., A.M.O., A.Z., J.M., U.B., N.K.S., J.K., V.K.Y. and J.Y. performed the LCLS experiment; I.D.Y., A.S.B., N.W.M., J.M.H., P.D.A. and N.K.S. developed new software for data processing; I.D.Y., A.S.B., L.L., F.D.F., C.W., T.F., P.A., H.D., J.M.H., N.K.S. and J.K. processed and analysed XFEL data; J.M., J.K., V.K.Y. and J.Y. wrote the manuscript with input from all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0681-2>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0681-2>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to A.Z., J.M., J.Y. or V.K.Y.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Sample preparation. PSII dimers were extracted and purified from *Thermosynechococcus elongatus* as reported previously³². PSII crystals ranging in size from 20 to 60 μm were then prepared using a modified seeding protocol³³. The crystals were dehydrated by treatment with high concentrations of PEG 5000. The final crystal suspension used for XRD measurements was in 0.1 M MES pH 6.5, 0.1 M ammonium chloride and 35% (w/v) PEG 5000, with \sim 0.5–0.8 mM chlorophyll concentration. After loading into the sample delivery syringe (Hamilton gastight syringe, 1,000 μl), each sample syringe was given a preflash using green LED diodes (525 nm, Thorlabs, USA) to synchronize the samples, ensure that all centres have an oxidized tyrosine D and maximize yield of the higher S states upon subsequent flashing. The photon flux used at the sample for the preflash was $2 \mu\text{mol m}^{-2} \text{s}^{-1}$. The samples were exposed for ten seconds while rotating the syringe, and dark-adapted for 30 min to 2 h after the preflash. We note that the PSII core complexes in our sample preparation contain a sufficient number of natural quinones to drive the catalytic reaction through the cycle^{8,34,35}.

Characterization. O_2 activity. Measurement of the O_2 yield by means of a Clark-type electrode under continuous illumination showed that the O_2 evolution rate of PSII solution before crystallization was $2,300 \pm 100 \mu\text{mol O}_2/(\text{mg}(\text{Chl}) \times \text{h})$ and after crystallization in the final buffer comprising 0.1 M MES pH 6.5, 0.1 M ammonium chloride it was $2,100 \pm 80 \mu\text{mol O}_2/(\text{mg}(\text{Chl}) \times \text{h})$ with 0.4 mM PPBQ. Approximately 90% activity was retained after crystallization.

EPR measurements. All sample batches used at the LCLS were checked with EPR for S-state turnover and Mn(II) content. The low-temperature X-band EPR spectra were measured using a Varian E109 EPR spectrometer equipped with a Model 102 Microwave bridge. For the turnover measurements, sample temperature was maintained at 8 K using an Air Products LTR liquid helium cryostat. The following spectrometer conditions were used: microwave frequency, 9.22 GHz; field modulation amplitude, 32 G at 100 kHz; microwave power, 20 mW. The turnover in the crystals was characterized using EPR spectroscopy. The turnover of the samples was measured as a function of the multiline EPR signal of the S_2 state, which oscillates with a period of four as a function of flash number. We observed the EPR signal for samples after a single turnover flash (1F) and monitored turnover by following decreases in amplitude of the multiline signal by subsequent flashes. In order to check the Mn(II) content, sample temperature was maintained at 20 K and the sensitivity of the measurement allowed to detect the presence of as low as 2% Mn(II) (compared to total Mn content) in the sample. The following spectrometer conditions were used: microwave frequency, 9.22 GHz; field modulation amplitude, 32 G at 100 kHz; microwave power, 1 mW. The Mn (II) content estimated by EPR agreed with that determined by in situ XES measurements³⁶.

Membrane inlet mass spectroscopy measurements. The S-state advancement of crystals was also evaluated by membrane inlet mass spectroscopy (MIMS). A crystal suspension with approximately 10% ^{18}O -labelled water was placed in the thin-layer MIMS setup and subjected to a laser preflash before dark adaptation for 40 min at room temperature. The sample was then subjected to 2F at 5 Hz frequency and the O_2 yield was detected as a peak at $m/z = 34$. The procedure was repeated for 3F and 4F using new crystal suspensions each time. The O_2 yield pattern as function of flash number was calculated by subtracting the normalized O_2 yield of a flash number with the yield from the preceding flash number and normalized to 3F. The O_2 pattern can be fitted satisfactorily with an average miss parameter of 22%. The estimated S-state population is presented in Extended Data Fig. 2.

Determination of the S-state population. We evaluated the S-state advancement of PSII crystal suspension by two methods, in situ and ex situ. X-ray emission spectroscopy, which monitors the oxidation state of Mn, was collected simultaneously with the XRD data (Fig. 1c) as described previously^{10,36,37}. The obtained single crystal XES spectra and details about XES data evaluation have been published recently³⁶. Standard deviations for the first moments of the XES data shown as error bars in Fig. 1c were determined by random sampling of each of the data sets 1,000 times (each time randomly splitting the data into two subsets) and then calculating the standard deviation from the resulting 2,000 spectra for each flash state³⁸.

Flash-induced oxygen measurements using MIMS for O_2 detection^{38–40} were carried out before the XFEL experiment. We estimated the S-state population from both methods and these are presented in Extended Data Fig. 2. For the purpose of refinement of the structural models in the 2F data, we fixed the S_3 populations to 70% and S_2 to 30%. For the 3F data, we fixed the S_3 populations to 40%, and fit the remaining 60% as S_0 , because S_0 is the major population within the 60%. Further details regarding the refinement of the mixed models in the 2F and 3F data are given below (Model building and map calculation).

Sample injection and illumination. The crystallography data were collected at the MXF instrument of LCLS^{41,42} during experiments LN84 and LQ39. The drop-on-tape (DOT) sample delivery method was used in combination with acoustic droplet ejection (ADE)³⁷. For capturing the stable intermediates S_2 , S_3 , and S_0 , each droplet of the crystal suspension was illuminated by 120-ns laser pulses at 527 nm using a Nd:YLF laser (Evolution, Coherent) via fibre-coupled outputs 1, 2 and/or 3,

resulting in a delay time of 0.2 s between each illumination, and of 0.2 s between the last illumination and the X-ray probe³⁷. We implemented a feedback control system of the belt speed and deposition delay, and the flashing delay and droplet phase were adjusted accordingly³⁷.

For ex situ testing of light saturation for the DOT system, a 100–150 μm thick sample film was established with the help of a washer between the silicon membrane of the mass spectrometer inlet and a thin microscope glass plate (thin layer MIMS setup). In this experiment, the samples were saturated at 70 mJ/cm^2 . The details have been described⁸. At the XFEL, a light intensity of $120 \pm 10 \text{ mJ/cm}^2$ was applied.

X-ray diffraction setup and data processing. PSII crystals were measured using X-ray pulses of \sim 40 fs length at 9.5 keV and with an X-ray spot size at the sample of \sim 3 μm in diameter. XRD data were collected on a Rayonix MX170 HS detector operating in the 2-by-2 binning mode at its maximum frame rate of 10 Hz. This mode provided the optimal trade-off of resolving power between adjacent Bragg reflections and quantity of images collected.

We developed the cctbx.xfel graphical user interface to track diffraction data acquisition, provide real-time feedback, and submit processing jobs. Processing jobs used dials.stills_process, a program within the cctbx.xfel framework that carries out lattice indexing, crystal model refinement, and integration and adopts a variety of defaults suited to XFEL still images^{43–48}. For each image, strong spots are first selected. Next, candidate basis vectors describing the lattice of strong spots are identified, and an optional target cell is used to filter these candidates. A crystal model (composed of a unit cell and crystal orientation) is then refined to minimize differences between observed spot centroids and predicted positions, and this model is used to generate a complete set of indexed positions on the frame. Finally, signal at these positions is integrated and any corrections or uncertainties are taken into account. We found that with the stills-specific defaults and very few non-default parameters, 20–50% of shots (which we estimate to be the majority of the shots containing crystals) could be successfully indexed.

The powder diffraction pattern of a silver(I) behenate sample (Alfa Aesar) in a quartz capillary (Hampton Research, 10 μm wall thickness) was used to obtain an initial estimate of detector distance. Initial indexing results were used to refine a detector distance and position for each interval between adjustments to the sample delivery system or detector position. These higher-precision detector positions were used in subsequent indexing and integration trials, resulting in a maximum of four distinct lattices indexed on a single shot.

Cluster.unit_cell, a command line tool in cctbx that clusters similar unit cells according to the Andrews–Bernstein distance metric^{49,50}, was used to obtain the average unit cell. This unit cell was used as the target unit cell when reprocessing all experimental data with dials.stills_process.

A total of 1,565,863 integrated lattices were obtained using dials.stills_process with a target unit cell of $a = 117.5 \text{ \AA}$, $b = 222.8 \text{ \AA}$, $c = 309.6 \text{ \AA}$, $\alpha = \beta = \gamma = 90^\circ$ and the space group $P2_12_12_1$. Signal was integrated to the edges of the detector in anticipation of a per-image resolution cutoff during the merging step. Integrated intensities were corrected for absorption by the Kapton conveyor belt to match the position of the belt and crystals relative to the X-ray beam³⁷.

Finally, XES data collected simultaneously with the diffraction images were used to sort out and exclude any sample batches that indicated the presence of Mn(II) released during the on-site crystallization³⁶. It was also used to confirm the advancement of S states by fibre-coupled lasers and a free space laser.

Image sets were also culled to include only images diffracting beyond 6 \AA (for small data sets) or beyond 3 \AA (all others), similar to a procedure that has been used previously⁵¹ to improve statistics in large data sets suffering from contamination by low-quality images. Until the experiments described here, we were typically data-limited and have focused data processing methods development on discovering how to extract the most signal from low-multiplicity data sets^{52,53}. Several data visualization tools we implemented in the cctbx.xfel graphical user interface have made it possible to tune crystallization conditions and the sample delivery system to optimize diffraction quality early in an XFEL diffraction experiment, resulting in collection of much larger quantities of data. Although post-refinement and the per-image resolution cut-offs used in cxi.merge downweight or remove most spot predictions without signal, we still observed improvement in merging statistics and map quality when excluding lattices diffracting to a resolution poorer than 3 \AA from these larger data sets, probably because of the limitation in orientational precision when indexing the small number of reflections visible on low-resolution stills.

The remaining integrated images were merged using cxi.merge as described previously⁸, with a couple of modifications. The default unit cell outlier rejection mechanism in cxi.merge was sufficiently selective on the image set curated as described above, so a pre-filtering step was not necessary. Also, a reference model and data set with a compatible unit cell—used by cxi.merge during scaling—were available from previous beam times, so a preliminary merging step with PRIME was not necessary.

Final merged data sets were acquired for the 0F, 1F, 2F (150 μ s), 2F (400 μ s), 2F, and 3F states to resolutions between 2.50 and 2.04 \AA , containing between 4,231 and 30,366 images (Extended Data Table 1). Additionally, data from all illuminated states were aggregated, culled to the subset of images extending past 2.2 \AA , and merged as a separate 'combined' data set to 1.98 \AA (results not shown).

Model building and map calculation. Initial structure refinement against the 'combined' data set at 1.98 \AA was carried out starting from a previously acquired high-resolution PSII structure in the same unit cell (PDB ID: 5TIS) using phenix.refine^{54,55}. After an initial rigid body refinement step, *xyz* coordinates and isotropic *B*-factors were refined for tens of cycles with automatic water placement enabled. Custom bonding restraints were used for the OEC (with large sigma values, to reduce the effect of the strain at the OEC on the coordinate refinement), chlorophyll-*a* (CLA, to allow correct placement of the Mg relative to the plane of the porphyrin ring), and unknown lipid-like ligands (STE). Custom coordination restraints overrode van der Waals repulsion for coordinated chlorophyll Mg atoms, the non-haem iron, and the OEC. Following real space refinement in Coot⁵⁶ of selected individual sidechains and the PsbO loop region and placement of additional water molecules, the model was refined for several additional cycles with occupancy refinement enabled, then as before without automatic water placement, and then as before with hydrogen atoms. NHQ flips and automatic linking were disabled throughout. A final 'combined' data set model was obtained with $R_{\text{work}}/R_{\text{free}}$ of 17.92%/22.01%.

The above model was subsequently refined against the illuminated data sets to produce models that differed primarily at the OEC and plastoquinone, as confirmed by isomorphous difference maps, with the lattermost refinement settings and different OEC bonding restraints. OEC bonding restraints for the 0F data set prevented large deviations from the high-resolution dark state OEC structure reported by Suga et al. (PDB ID: 4UB6)⁵³. Bonding restraints for the other data sets loosely restrained the models to metal–metal distances matching spectroscopic data and metal–oxygen distances matching the most likely proposed models^{57–61}. A number of ordered water positions were excluded from subsequent automatic water placement rounds by renaming the residue names to OOO and supplying Phenix with a bonding restraint CIF dictionary for OOO identical to that for HOH, and the waters coordinating the OEC were incorporated into the OEC restraint CIF file directly. After 12–15 of cycles of refinement in this manner, individual illuminated states at various resolutions were obtained ranging in $R_{\text{work}}/R_{\text{free}}$ from 16.69%/24.60% to 19.33%/26.39% (Extended Data Table 1).

After the first cycles of refinement for the 2F data using the initial OEC model from the 0F data as starting point, a positive peak in the $mF_{\text{obs}} - DF_{\text{calc}}$ density close to Mn1 became visible (see also Extended Data Fig. 5) and the automatic water placement step in refinement placed a water at the OEC between O5 and Mn1. We designated this new water as a potential Ox and added it to the OEC. We tested several starting positions for this Ox, including a position similar to O6 (Suga et al.)⁹ at 1.5 \AA from O5, but refinement resulted in shifts of the Ox away from O5 and close to Mn1. $mF_{\text{obs}} - DF_{\text{calc}}$ difference maps calculated for different positions of this additional oxygen also confirmed a placement at about 1.8 \AA from Mn1 and 2.1 \AA from O5. For the final refinement, Ox was included in the CIF restraints for the OEC in the S₃ state.

To best approximate the contributions of dimers that did not advance to the next S state owing to illumination misses, for the 2F and 3F data sets, the 2F and 3F models were split into A and B alternate conformers in regions of chains A/a, C/c and D/d surrounding (and including) the OEC. These residues are A55–65, A160–190, A328–344, C328, C354–358 and D352. Population of the S₃ and S₀ states in the 2F and 3F data was estimated on the basis of oxygen evolution and XES measurements (Extended Data Fig. 2) and rounded to the nearest 10%, yielding 70% S₃ state population in the 2F data set and 60% S₀ state population in the 3F data set. Accordingly, for the 2F data set, the main conformer across this entire region was set at 0.7 occupancy, and the minor conformer was set at 0.3 occupancy. Analogously, for the 3F data set, conformers were set at 0.6 and 0.4 to match an estimated 40% contribution from the S₃ state and modelling the remaining 60% as the S₀ state. The major conformer was allowed to refine as usual, while the minor conformer was fixed during refinement and set to match the major conformer of the previous S state (for example, fixed coordinates for S₂ at 30% for the 2F, S₃-enriched state). This was achieved by least-squares fitting the refined model of the previous S state onto the new model at the split region in PyMol⁶² and replacing the minor conformer atomic coordinates with the fitted model coordinates, then excluding the newly placed atoms from refinement in phenix.

Although phenix.refine supports modelling of three or more conformers, we limited our analysis to two conformers in consideration of both the limits of the resolution and the precision of the S-state contribution estimates, and we did not model a ~10% contribution of the S₁ state in the 1F data set. When placing the refined S₃ state model into the 3F model at 0.4 occupancy, we used the 0.7 occupancy model refined as described above, not the combination of both conformers. This analysis was not possible for the 2F (150 μ s) and 2F (400 μ s) models because

the complementary components would be time points 150 μ s and 400 μ s after the S₁ state, structures we have not probed.

Estimated positional precision. The maximum-likelihood coordinate error calculated during refinement is a general-purpose metric for positional error but is subject to several limitations, including the impact of bonding, angle, coordination and other restraints on the refined model. Previously, we have generated ballpark estimates of positional error for various sections of the model by setting them to zero occupancy, conducting simulated annealing followed by refinement, placing the same components into omit density, and reporting the final magnitude of the shift between the centres of mass of the original and omit density-fitted components⁸. This is impractical for more than a handful of representative cofactors or segments of the main chain and is difficult for much smaller groups of atoms or individual atoms whose environments easily fill the missing density region if it is not artificially held open. We also tried multi-start kicked model refinement and found that the OEC refined back to nearly the same position in all trials. We therefore shifted our focus to a tool that perturbs the structure factors directly. By perturbing structure factors by $\pm |F_{\text{obs}} - F_{\text{model}}|$ in 100 trials using the END/RAPID command line tools, we added noise proportional to the error in the model to generate 100 perturbed data sets for each illumination state, re-refined kicked models against each new data set, and calculated the mean and s.d. of selected bond distances across the re-refined models⁶³. Metal–metal distances at the OEC had standard deviations between 0.08 and 0.12 \AA across these trials, while distances between OEC metals and bridging oxygen atoms had standard deviations varying between 0.10 and 0.23 \AA and distances between OEC metals and coordinating ligands were found to have standard deviations between 0.13 and 0.20 \AA .

Estimating the uncertainty of the omit densities. Changes in the electron density at the positions of Ox and O5 were obtained from O5 and Ox omit maps and normalized against the average electron density maximum at the O2 position in O2 omit maps, assuming that O2 is always fully occupied in the different flash states. The standard deviation of the electron density value at O2 over all data sets and both monomers was also used to estimate the uncertainty of the normalized omit density. The omit densities of a particular data set were divided by the average omit densities of the chloride ions of the same data set to equate the densities from different data sets.

Isomorphous difference maps. Slight, nonphysical differences in merged unit cells were modelled across the illumination states in this sequence of data sets. Large distributions of unit cells derived from indexing with dials.stills_process are known to reflect uncertainty in the crystal model, not variation among actual crystals⁶⁴, and the distributions also shifted as sample-to-detector distance changed over the course of an LCLS shift. Because it was not possible to cycle through all illumination conditions throughout each of the experiments, the average unit cell dimensions varied across data sets as well, resulting in the aforementioned non-physical differences in merged unit cells. To obtain isomorphous difference maps without artefacts from these apparent unit cell variations we computed a second set of data, selecting only lattices with unit cells within 1% of the target unit cell. The resulting smaller data sets were at slightly reduced resolution (Extended Data Table 1b). However, because the merged unit cell differences were small, artefact free isomorphous difference maps could be calculated between pairs of these data sets. The corresponding .mtz and .pdb files for these smaller data sets are available from the authors upon request.

Code availability. The open source programs dials.stills_process, the cctbx.xfel GUI and cxi.merge are distributed with DIALS packages available at <http://dials.github.io>, with further documentation available at <http://cci.lbl.gov/xfel>.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The atomic coordinates and structure factors have been deposited in the Protein Data Bank under the following accession codes: 6DHE for the 0F, 6DHF for the 1F, 6DHO for the 2F, 6DHG for the 2F(150 μ s), 6DHH for the 2F(400 μ s) and 6DHP for the 3F data.

32. Hellmich, J. et al. Native-like photosystem II superstructure at 2.44 \AA resolution through detergent extraction from the protein crystal. *Structure* **22**, 1607–1615 (2014).
33. Ibrahim, M. et al. Improvements in serial femtosecond crystallography of photosystem II by optimizing crystal uniformity using microseeding procedures. *Struct. Dyn.* **2**, 041705 (2015).
34. Guskov, A. et al. Cyanobacterial photosystem II at 2.9- \AA resolution and the role of quinones, lipids, channels and chloride. *Nat. Struct. Mol. Biol.* **16**, 334–342 (2009).
35. Krivanek, R., Kern, J., Zouni, A., Dau, H. & Haumann, M. Spare quinones in the Q_B cavity of crystallized photosystem II from *Thermosynechococcus elongatus*. *Biochim. Biophys. Acta* **1767**, 520–527 (2007).

36. Fransson, T. et al. X-ray emission spectroscopy as an *in situ* diagnostic tool for X-ray crystallography of metalloproteins using an X-ray free-electron laser. *Biochemistry* **57**, 4629–4637 (2018).

37. Fuller, F. D. et al. Drop-on-demand sample delivery for studying biocatalysts in action at X-ray free-electron lasers. *Nat. Methods* **14**, 443–449 (2017).

38. Kern, J. et al. Taking snapshots of photosynthetic water oxidation using femtosecond X-ray diffraction and spectroscopy. *Nat. Commun.* **5**, 4371 (2014).

39. Yano, J. et al. in *Sustaining Life on Planet Earth: Metalloenzymes Mastering Dioxygen and Other Chewy Gases, Metal Ions in Life Sciences* (eds Kroneck, P. M. H. & Sosa Torres, M. E.) 13–43 (Springer International Publishing, 2015).

40. Beckmann, K., Messinger, J., Badger, M. R., Wydrzynski, T. & Hillier, W. On-line mass spectrometry: membrane inlet sampling. *Photosynth. Res.* **102**, 511–522 (2009).

41. Emma, P. et al. First lasing and operation of an Ångström-wavelength free-electron laser. *Nat. Photon.* **4**, 641–647 (2010).

42. Boutet, S., Cohen, A. & Wakatsuki, S. The new macromolecular femtosecond crystallography (MFX) instrument at LCLS. *Synchr. Radiat. News* **29**, 23–28 (2016).

43. Sauter, N. K. XFEL diffraction: developing processing methods to optimize data quality. *J. Synchr. Radiat.* **22**, 239–248 (2015).

44. Winter, G. et al. DIALS: implementation and evaluation of a new integration package. *Acta Crystallogr. D Struct. Biol.* **74**, 85–97 (2018).

45. Sauter, N. K., Hattne, J., Grosse-Kunstleve, R. W. & Echols, N. New Python-based methods for data processing. *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1274–1282 (2013).

46. Hattne, J. et al. Accurate macromolecular structures using minimal measurements from X-ray free-electron lasers. *Nat. Methods* **11**, 545–548 (2014).

47. Sauter, N. K. et al. Improved crystal orientation and physical properties from single-shot XFEL stills. *Acta Crystallogr. D Biol. Crystallogr.* **70**, 3299–3309 (2014).

48. Waterman, D. G. et al. Diffraction-geometry refinement in the DIALS framework. *Acta Crystallogr. D Struct. Biol.* **72**, 558–575 (2016).

49. Zeldin, O. B. et al. Data Exploration Toolkit for serial diffraction experiments. *Acta Crystallogr. D Biol. Crystallogr.* **71**, 352–356 (2015).

50. Andrews, L. C. & Bernstein, H. J. The geometry of Niggli reduction: BGAOL - embedding Niggli reduction and analysis of boundaries. *J. Appl. Crystallogr.* **47**, 346–359 (2014).

51. Suga, M. et al. Native structure of photosystem II at 1.95 Å resolution viewed by femtosecond X-ray pulses. *Nature* **517**, 99–103 (2015).

52. Uervirojanaengkorn, M. et al. Enabling X-ray free electron laser crystallography for challenging biological systems from a limited number of crystals. *eLife* **4**, e05421 (2015).

53. Lyubimov, A. Y. et al. Advances in X-ray free electron laser (XFEL) diffraction data processing applied to the crystal structure of the synaptotagmin-1/SNARE complex. *eLife* **5**, e18740 (2016).

54. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).

55. Afonine, P. V. et al. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 352–367 (2012).

56. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).

57. Wieghardt, K. The active-sites in manganese-containing metalloproteins and inorganic model complexes. *Angew. Chem. Int. Edn Engl.* **28**, 1153–1172 (1989).

58. Cinco, R. M. et al. Comparison of the manganese cluster in oxygen-evolving photosystem II with distorted cubane manganese compounds through X-ray absorption spectroscopy. *Inorg. Chem.* **38**, 5988–5998 (1999).

59. Mukhopadhyay, S., Mandal, S. K., Bhaduri, S. & Armstrong, W. H. Manganese clusters with relevance to photosystem II. *Chem. Rev.* **104**, 3981–4026 (2004).

60. Law, N. A., Caudle, M. T. & Pecoraro, V. L. Manganese redox enzymes and model systems: Properties, structures, and reactivity. *Adv. Inorg. Chem.* **46**, 305–440 (1999).

61. Tsui, E. Y., Kanady, J. S. & Agapie, T. Synthetic cluster models of biological and heterogeneous manganese catalysts for O₂ evolution. *Inorg. Chem.* **52**, 13833–13848 (2013).

62. Schrödinger LLC. The pymol molecular graphics system, version 1.8. (2015).

63. Lang, P. T., Holton, J. M., Fraser, J. S. & Alber, T. Protein structural ensembles are revealed by redefining X-ray electron density noise. *Proc. Natl. Acad. Sci. USA* **111**, 237–242 (2014).

64. Brewster, A. S. et al. Improving signal strength in serial crystallography with DIALS geometry refinement. *Acta Crystallogr. D Struct. Biol.* **74**, 877–894 (2018).

65. Kupitz, C. et al. Serial time-resolved crystallography of photosystem II using a femtosecond X-ray laser. *Nature* **513**, 261–265 (2014).

66. Ho, F. M. & Styring, S. Access channels and methanol binding site to the CaMn₄ cluster in Photosystem II based on solvent accessibility simulations, with implications for substrate water access. *Biochim. Biophys. Acta* **1777**, 140–153 (2008).

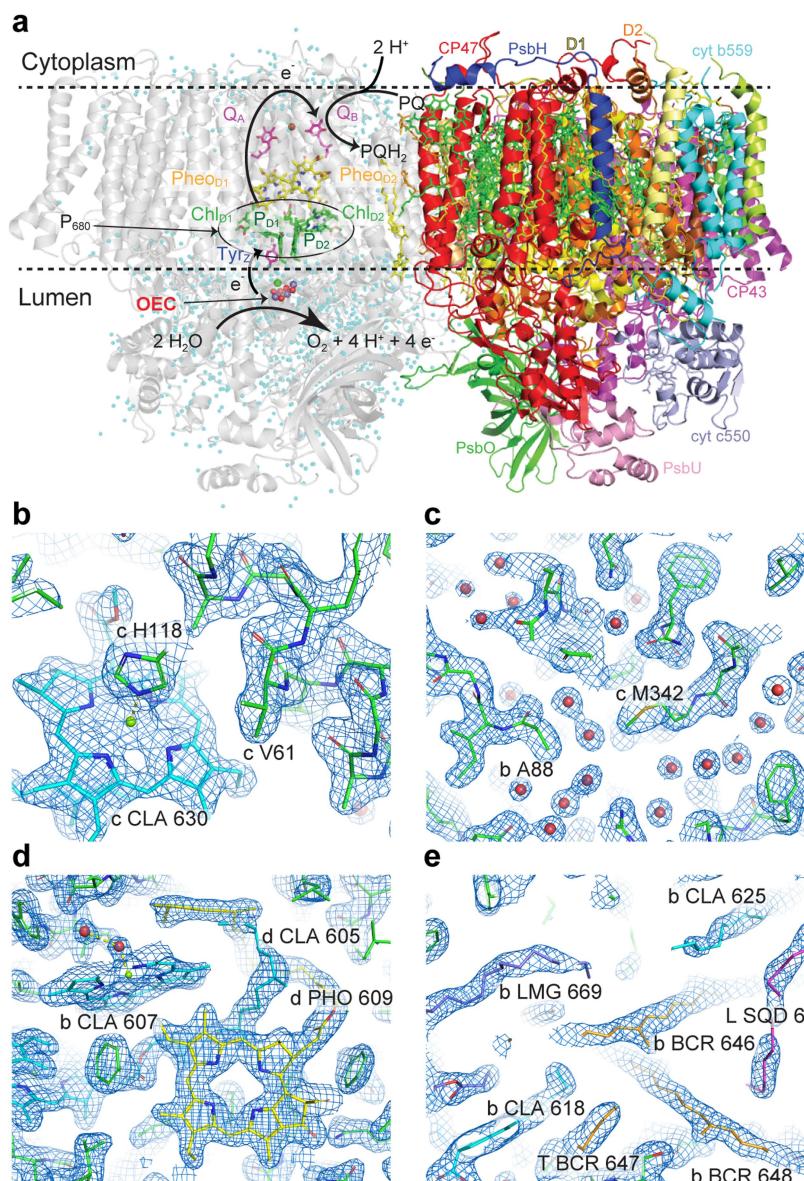
67. Vassiliev, S., Zaraiskaya, T. & Bruce, D. Exploring the energetics of water permeation in photosystem II by multiple steered molecular dynamics simulations. *Biochim. Biophys. Acta* **1817**, 1671–1678 (2012).

68. Murray, J. W. & Barber, J. Structural characteristics of channels and pathways in photosystem II including the identification of an oxygen channel. *J. Struct. Biol.* **159**, 228–237 (2007).

69. Gabdulkhakov, A. et al. Probing the accessibility of the Mn₄Ca cluster in photosystem II: channels calculation, noble gas derivatization, and cocrystallization with DMSO. *Structure* **17**, 1223–1234 (2009).

70. Umena, Y., Kawakami, K., Shen, J.-R. & Kamiya, N. Crystal structure of oxygen-evolving photosystem II at a resolution of 1.9 Å. *Nature* **473**, 55–60 (2011).

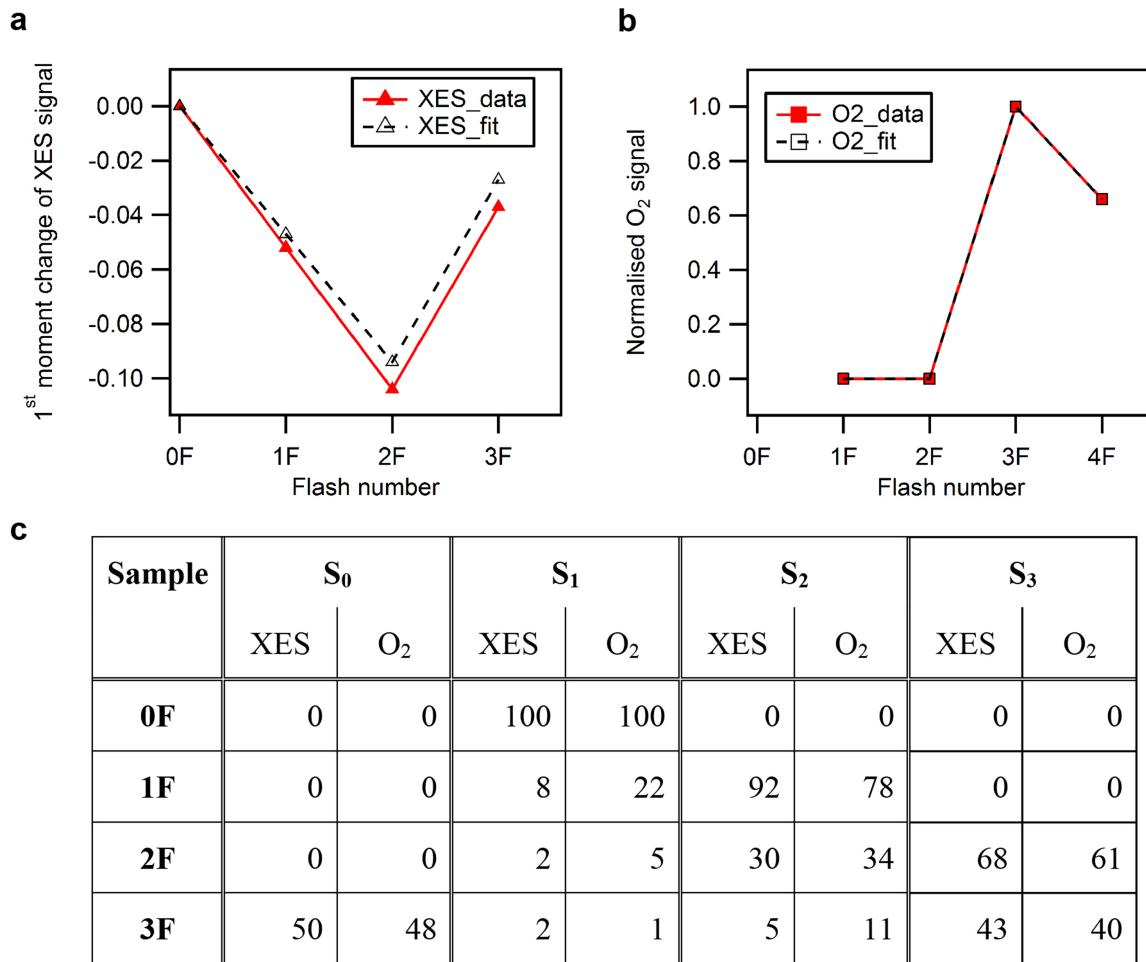
71. Sakashita, N., Watanabe, H. C., Ikeda, T. & Ishikita, H. Structurally conserved channels in cyanobacterial and plant photosystem II. *Photosynth. Res.* **133**, 75–85 (2017).



Extended Data Fig. 1 | Overview of the PSII structure and electron density maps of the 3F state. **a**, Structure of the native PSII homodimer. In the left monomer the location of cofactors for the initial charge separation (P_{680} , Pheo_{D1}), and for the electron transfer leading to the reduction of the plastoquinone (Q_A , Q_B) at the acceptor side and to the oxidation of the OEC at the donor side by $P680^+$ are indicated. In the right monomer, the

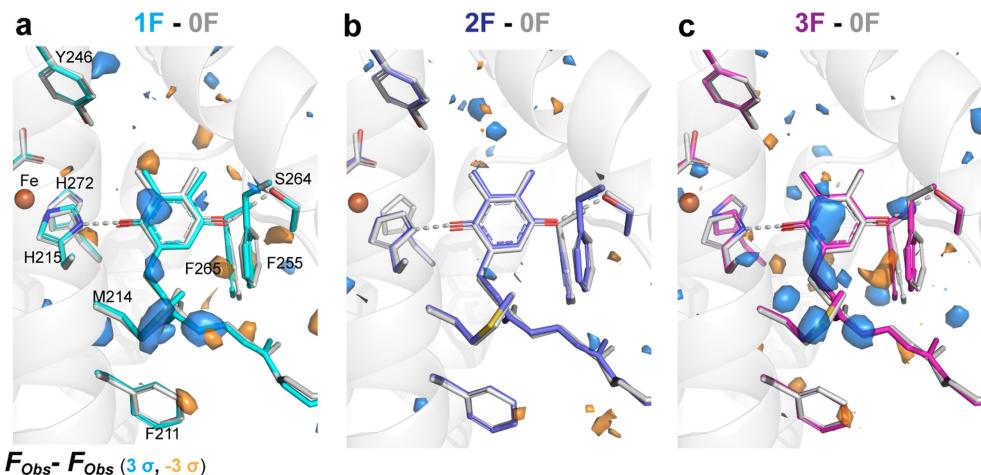
locations of the protein subunits are displayed. **b–d**, $2mF_{\text{obs}} - DF_{\text{calc}}$ map (blue, 1.5σ contour) obtained from the room temperature 3F data set.

b, Density around the main chain and a chlorophyll. **c**, Well-resolved ordered water molecules. **d**, Chlorophyll and pheophytin molecules with well-resolved tails. **e**, Clear density in hydrophobic regions and along cofactor hydrocarbon tails.



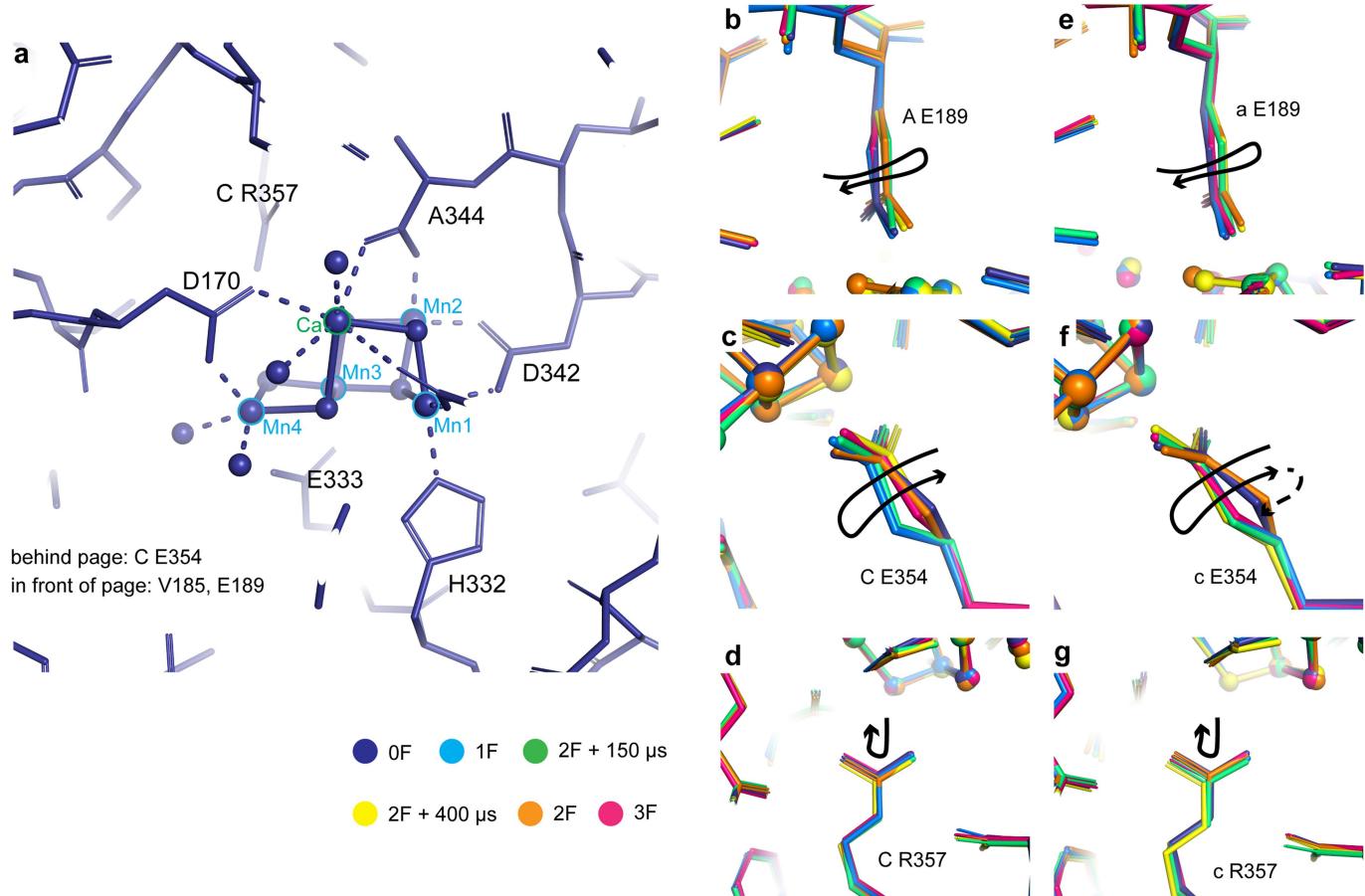
Extended Data Fig. 2 | Flash-induced S-state turnover of PSII microcrystals. **a**, Change of the first moment of the in situ-measured Mn K β XES as a function of flashes and fit to the data. **b**, Flash-induced O₂ yield as measured by MIMS as a function of flash number and fit to the data. **c**, The estimated S-state population (%) for each of the flash states from fitting of the XES data and of the flash-induced O₂ evolution pattern of a suspension of PSII crystals at pH 6.5. Two different fits were performed: a global fit

of both O₂ and XES data using an equal miss parameter of 22% and 100% S₁ population in the 0F sample (black traces in **a**, **b**; S-state distribution listed in the columns headed O₂), and a direct fit of the XES data using a 8% miss parameter in the S₁→S₂ and a 27% miss parameter for the S₂→S₃ and S₃→S₀ transitions (XES in **c**). For the XES fit, shifts of -0.06 eV per oxidation state increase for all S states were assumed. The XES raw spectra are published elsewhere³⁶.



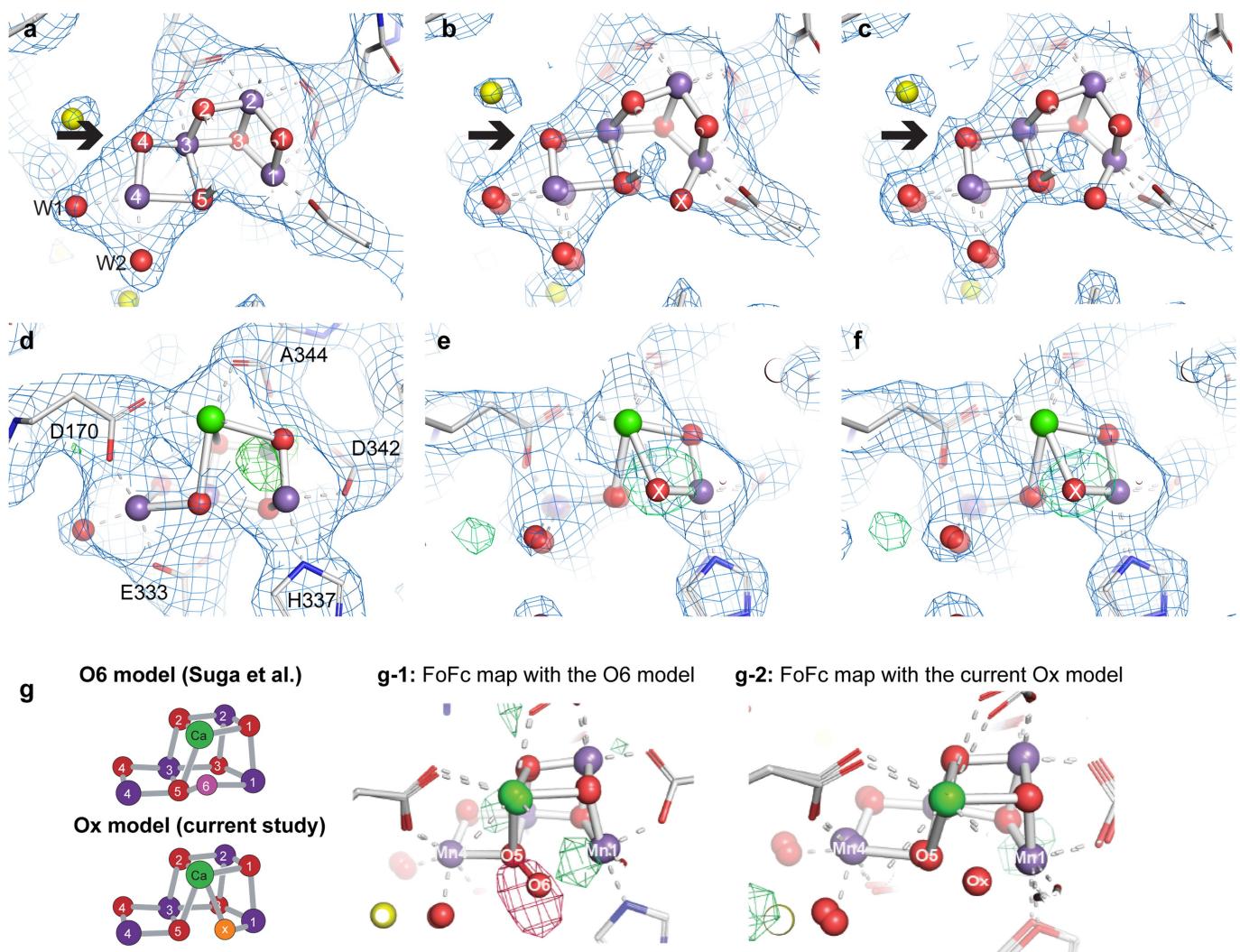
Extended Data Fig. 3 | Isomorphous difference maps in the second monomer at the Q_B site. **a–c**, $F_{\text{obs}} - F_{\text{obs}}$ maps contoured at 3σ at plastoquinone Q_B in monomer a. **a**, 1F – 0F difference map matching reduction of the plastoquinone to a semiquinone and concomitant slight geometry change. **b**, 2F – 0F difference map matching replacement of the

fully reduced quinol with another quinone at the original position. **c**, 3F – 0F difference map, showing again structural changes similar to the 1F – 0F map, indicating formation of the semiquinone. Similar views are shown for monomer A in Fig. 1d–f and comparison of both monomers indicates similar flash-induced changes in both monomers.



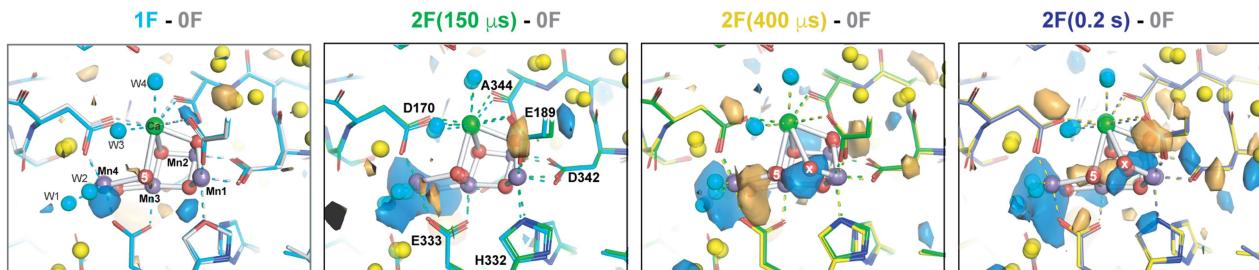
Extended Data Fig. 4 | Movement of ligands around the OEC in the different S states. **a**, Overview of the ligand environment of the OEC, showing the dark state (0F) structure. Coordination of the OEC by nearby side chains and water molecules is indicated by dashed lines. **b–d**, Trends for selected individual side chains in both monomers (**b–d**, monomer A; **e–g**, monomer a). Overlays of the refined models at the OEC following

least-squares fitting of subunit D1 residues 55–65, 160–190 and 328, subunit CP43 residues 328 and 354–358, and chain D residue 352 of each other model to the 0F model. The largest and most consistent motions of side chains near the OEC through the sequence of illuminated state models are annotated with arrows indicating the trend. A motion observed in only one monomer is indicated by a dashed line.



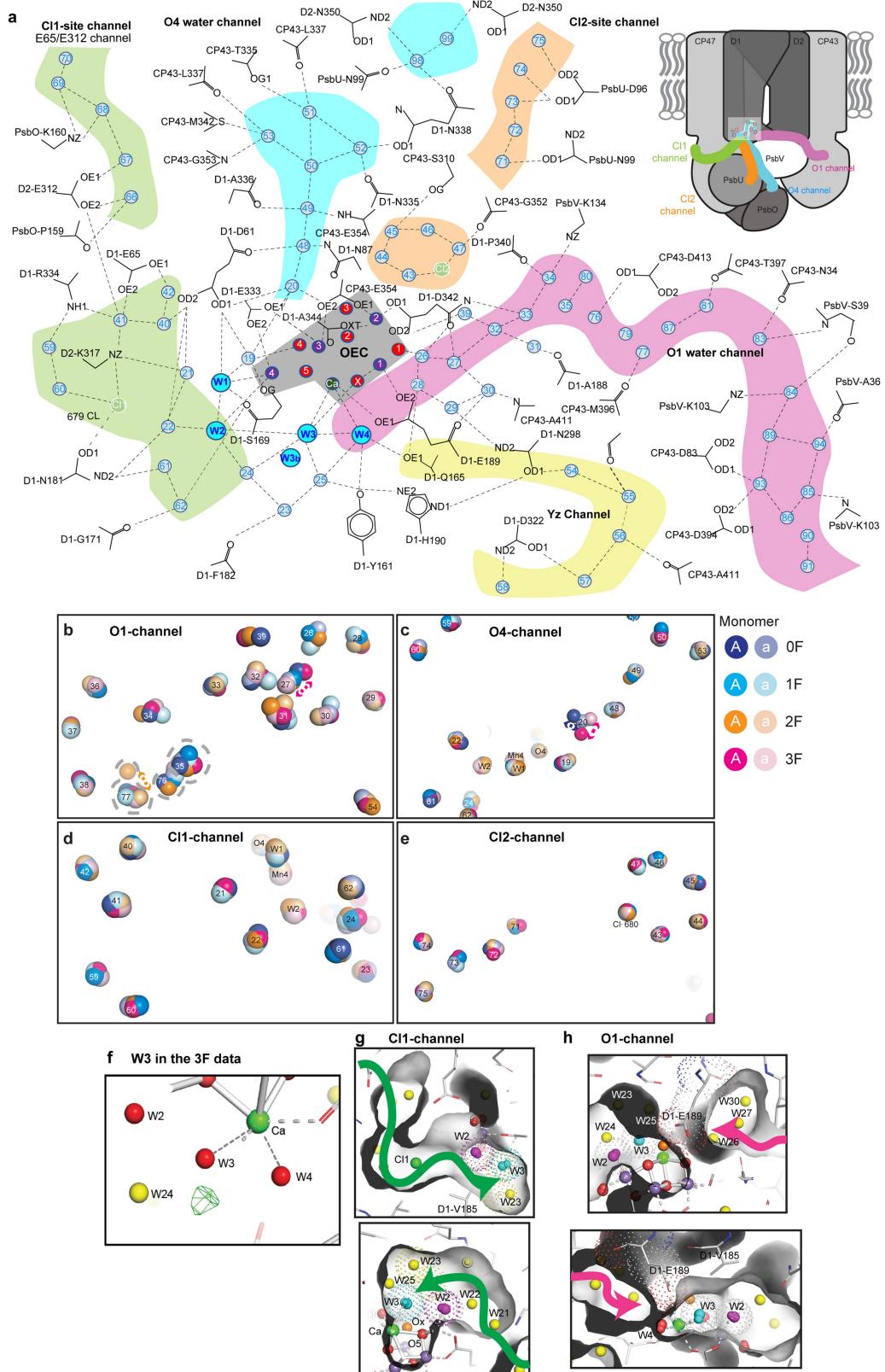
Extended Data Fig. 5 | Impact of the data quality on the resolving power of the maps. **a–f**, The data quality evidenced by 2F state models and $2mF_{\text{obs}} - DF_{\text{calc}}$ maps contoured at 1.5σ . **a**, 5TIS (2F, 2.25 Å) model and map. Overlays indicate atom numbering in the OEC and the identities of selected coordinating sidechains. **b**, Current 2F model and map cut to 2.25 Å. **c**, Current 2F model and map at the full 2.07 Å resolution. Emergence of locations of O4 with improved data quality is indicated by bold arrow. **d**, As in **a** from a different angle and with $mF_{\text{obs}} - DF_{\text{calc}}$ density at 3σ indicating the lack of sufficient evidence for inserting an additional O atom at a chemically reasonable position. **e**, **f**, As in **b**, **c** from the same direction as **d** and with $mF_{\text{obs}} - DF_{\text{calc}}$ density at 3σ shown to 2.25 and 2.05 Å, respectively, after omitting the inserted Ox

atom. Centring of the refined Ox position within the omit density gives a clear indication of the position of the inserted water in the S_3 state with the current, higher-quality data, even when artificially cut to the same resolution as the previous data set. **g**, $mF_{\text{obs}} - DF_{\text{calc}}$ maps of the 2F data that compare the O6 model from Suga et al.⁹ and the Ox model from the current study. Map shows the $mF_{\text{obs}} - DF_{\text{calc}}$ density calculated with our current 2F data and our model adding the O6 position of Suga et al.⁹ (with the occupancy of 0.7 and B-factor of 30) (**g-1**), and with our Ox model (**g-2**). We see clearly a positive density for the missing Ox and a negative density at the O6 position in **g-1**. Schematics of the O6 and Ox S_3 models are shown on the left.



Extended Data Fig. 6 | Isomorphous difference maps in the second monomer at the OEC. Isomorphous difference density OEC sites in monomer a. $F_{\text{obs}} - F_{\text{obs}}$ difference densities between the various illuminated states and the OF data are contoured at $+3\sigma$ (blue) and

-3σ (orange). The model for the OF data is shown in light grey whereas carbons are coloured as follows: 1F (cyan), 2F (150 μs) (green), 2F (400 μs) (yellow) and 2F (0.2 s) (blue).



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Water environment of the OEC. **a**, Extended schematic of the hydrogen bonding network connecting the OEC to the solvent-exposed surface of PSII and identification of several channels for either possible water movement or proton transfer. Top right, locations of four selected channels in the PSII monomer. **b–e**, Movements within the water networks across monomers. Coloured spheres are shown for each ordered water or chloride ion across the four metastable states, 0F through 3F, and for both monomers, with the stronger colour matching the first (A) monomer and the lighter colour matching the second (a) monomer. For ordered solvent, residue number is shown; for OEC atoms, the atom identifier is shown; and for the Cl2 site, the Cl^- 680 label is shown. **b**, The O1 water chain. Positional disagreement between monomers is visible especially near waters 77 (2F) and 27 (3F) and is on the same scale as changes between illuminated states, both of which may indicate a more dynamic water channel. **c**, The O4 water chain. With the notable exception of water 20, most water positions are stable across monomers

and illuminated states. Water 20 is highly unstable in position in the two states (0F, 3F) in which it is modelled, and there is not sufficient density in the remaining states to model a water 20 position. **d**, The Cl1 site water channel with no notable movements. **e**, The Cl2 site water channel with no notable movements. **f**, Indication of a split position of W3 in the S_0 state. $mF_{\text{obs}} - DF_{\text{calc}}$ difference density (green mesh) in the 3F state suggests an alternate position near W3 (W3b in Fig. 4d). **g, h**, Possible access to W3/Ox side from the Cl1 or the O1 channel. The surface of the protein is shown in grey to visualize the extent of the cavities around the OEC, and Van der Waals radii are indicated for selected residues or atoms by dotted spheres. Shown are two different views for each channel. The direction of the Cl1 channel is indicated by a green arrow and the O1 channel by a pink arrow. Water W2 is shown in purple, W3 in cyan and Ox in orange. Yellow spheres indicate other waters. Mn are shown in magenta, other bridging oxygens as red spheres.

Extended Data Table 1 | Merging and refinement statistics for (a) the refined data sets including all lattices or (b) for the additional data sets containing only lattices with unit cells within 1% of the target unit cell

	0F	1F	2F (150 μ s)	2F (400 μ s)	2F [‡]	3F [‡]
Resolution range refined (Å)	30.552 - 2.05	30.427 - 2.08	30.783 - 2.5	30.851 - 2.2	30.578 - 2.07	31.005 - 2.04
Resolution range upper bin (Å)	2.085 - 2.05	2.116 - 2.08	2.543 - 2.5	2.238 - 2.2	2.106 - 2.07	2.075 - 2.04
Wavelength (Å)	1.303	1.303	1.301	1.301	1.303	1.303
Space group	P2 ₁ 2 ₁ 2 ₁					
Unit cell parameters (Å)	a=116.9 b=221.4 c=308.7	a=116.8 b=220.9 c=307.0	a=117.6 b=222.7 c=309.1	a=117.7 b=222.6 c=308.5	a=117.0 b=221.5 c=308.3	a=116.7 b=221.2 c=307.6
Images merged	30366	23744	4231	13072	24481	25134
Unique reflections (upper bin)	508701 (25202)	487161 (24128)	281837 (13902)	412258 (20410)	494189 (24482)	516201 (25594)
Completeness (upper bin)	99.95% (99.81%)	99.95% (99.78%)	99.92% (99.96%)	99.94% (99.91%)	99.95% (99.87%)	99.95% (99.86%)
CC _{1/2}	98.7% (1.8%)	97.5% (1.3%)	93.8% (6.8%)	96.0% (0.6%)	98.2% (2.1%)	98.6% (0.8%)
Multiplicity (upper bin)	186.8 (9.0)	160.1 (8.4)	45.4 (9.2)	98.4 (9.8)	156.8 (9.6)	170.7 (9.3)
Pred. multiplicity [*]	458.3	383.1	88.2	272.3	375.1	413.0
(upper bin)	(360.4)	(301.8)	(66.6)	(209.9)	(291.9)	(318.6)
I/σ _{Ha14(I)} [†]	16.6 (0.5)	15.0 (0.5)	11.1 (1.0)	10.6 (0.7)	15.5 (0.6)	17.0 (0.6)
Wilson B-factor	26.8	27.1	35.5	27.4	27.3	27.1
R-factor	18.48%	18.90%	16.69%	19.33%	18.44%	18.64%
R-free	23.99%	24.56%	24.60%	26.39%	24.75%	24.85%
Number of atoms	103732	103728	103713	103719	105764	105761
Number non-hydrogen atoms	52203	52199	52188	52194	53286	53283
Ligands	186	186	186	186	188	188
Waters	2021	2017	2012	2016	2015	2010
Protein residues	5306	5306	5306	5306	5306	5306
RMS (bonds)	0.014	0.015	0.015	0.016	0.014	0.014
RMS (angles)	1.51	1.53	1.57	1.61	1.52	1.50
Ramachandran favored	96.5%	95.8%	94.8%	95.3%	96.2%	96.3%
Ramachandran outliers	0.31%	0.38%	0.56%	0.38%	0.34%	0.31%
Clashscore	5.4	6.5	6.8	6.3	6.6	6.8
Average B-factor	42.8	44.4	50.3	46.5	43.4	42.3

	0F	1F	2F (150 μ s)	2F (400 μ s)	2F [‡]	3F [‡]
Resolution range refined (Å)	30.05 - 2.07	30.04 - 2.13	29.70 - 2.60	29.85 - 2.30	30.18 - 2.09	30.05 - 2.05
Resolution range upper bin (Å)	2.11 - 2.07	2.17 - 2.13	2.65 - 2.60	2.34 - 2.30	2.13 - 2.09	2.09 - 2.05
Wavelength (Å)	1.301	1.301	1.301	1.301	1.301	1.301
Space group	P2 ₁ 2 ₁ 2 ₁					
a=117.6 b=222.8	a=117.6 b=222.8	a=117.6 b=222.8	a=117.6 b=222.8	a=117.6 b=222.8	a=117.6 b=222.8	a=117.6 b=222.8
Unit cell parameters (Å)	c=309.7	c=309.7	c=309.7	c=309.7	c=309.7	c=309.7
Images merged	15669	10757	2087	4576	13603	14019
Unique reflections (upper bin)	490658 (24364)	450643 (22354)	249070 (12358)	358519 (17699)	476797 (23606)	505050 (25047)
Completeness (upper bin)	99.96% (99.98%)	99.95% (99.96%)	99.91% (99.93%)	99.93% (99.93%)	99.96% (99.93%)	99.95% (99.92%)
CC _{1/2}	98.2% (2.1%)	95.6% (2.8%)	92.4% (6.8%)	96.1% (2.9%)	97.7% (0.9%)	98.0% (0.4%)
Multiplicity (upper bin)	98.5 (11.1)	71.7 (10.6)	28.3 (9.8)	48.3 (10.4)	89.9 (11.0)	101.2 (10.6)
Pred. multiplicity [*]	204.8	144.2	41.3	86.3	187.6	210.1
(upper bin)	(153.1)	(107.2)	(29.81)	(62.32)	(139.22)	(154.24)
I/σ _{Ha14(I)} [†]	13.3 (0.7)	11.1 (0.7)	11.5 (1.1)	10.6 (0.8)	12.7 (0.7)	14.3 (0.6)
Wilson B-factor	24.4	24.4	31.8	24.0	24.8	24.9
R-factor	19.18%	19.75%	17.99%	19.82%	19.48%	19.44%
R-free	23.20%	24.43%	23.97%	24.94%	23.93%	23.62%
Number of atoms	52325	52154	50943	51500	52251	52319
Number non-hydrogen atoms	52325	52154	50943	51500	52251	52319
Ligands	187	187	187	187	187	187
Waters	2226	2054	843	1400	2151	2219
Protein residues	5300	5300	5300	5300	5300	5300
RMS (bonds)	0.008	0.009	0.009	0.009	0.008	0.008
RMS (angles)	0.98	1.06	1.08	1.07	0.97	0.97
Ramachandran favored	97.3%	97.1%	96.5%	96.8%	97.6%	97.5%
Ramachandran outliers	0.36%	0.31%	0.38%	0.27%	0.35%	0.27%
Clashscore	5.7	6.7	7.7	7.3	6.0	5.9
Average B-factor	37.6	39.9	44.2	44.4	38.9	38.6

*Predictions multiplicity is the multiplicity of all spot predictions matching the indexing solution on a given image, before a per-image resolution cutoff. Multiplicities for data sets merged by the Monte Carlo method (for example, Kupitz et al.⁶⁵, and Suga et al.⁹) without per-image resolution cutoffs are best compared with this metric.

[†]I/σ(I) calculation as described^{46,64}.

[‡]For the 2F and 3F data sets, a region of 66 amino acids around the OEC was modelled as double conformers to reflect the contribution from the two main S-state species in each of the data sets with contributions of 30% and 70% for S₂ and S₃ in the 2F and 40% and 60% for S₃ and S₀, respectively, in the 3F data sets.

Extended Data Table 2 | Interatomic distances at the OEC in each merged data set, in each monomer (A/a), in Å

Distances (Å)	0F		1F		2F (150 μs)		2F (400 μs)		2F OEC (S ₃)		3F OEC (S ₀)		
	A	a	A	a	A	a	A	a	A	a	A	a	
CA1-	MN1	3.41	3.45	3.46	3.38	3.49	3.45	3.47	3.58	3.37	3.37	3.44	3.31
	MN2	3.38	3.38	3.40	3.42	3.40	3.44	3.41	3.39	3.27	3.38	3.45	3.41
	MN3	3.52	3.49	3.54	3.50	3.42	3.53	3.49	3.57	3.52	3.61	3.55	3.52
	MN4	3.77	3.88	3.90	3.90	3.72	3.96	3.84	4.07	3.90	4.11	3.94	4.07
MN1-	MN2	2.77	2.77	2.85	2.77	2.89	2.78	2.86	2.72	2.77	2.73	2.81	2.71
	MN3	3.22	3.26	3.29	3.22	3.38	3.32	3.39	3.35	3.34	3.32	3.30	3.24
	MN4	4.80	4.91	4.84	4.88	5.01	5.05	5.05	5.16	5.01	5.11	4.97	4.98
MN2-	MN3	2.87	2.83	2.86	2.82	2.85	2.83	2.88	2.82	2.85	2.86	2.90	2.82
MN3-	MN4	2.70	2.77	2.70	2.79	2.70	2.84	2.74	2.90	2.70	2.84	2.79	2.91
MN1-	Ox						1.82	1.67	1.78	1.80			
MN4-	O5						2.23	2.28	2.17	2.27			
O5-	Ox						1.84*	1.87*	2.10	2.07			
MN4-	W1	2.18	2.13	2.06	2.17	2.03	1.97	2.37	2.23	2.16	2.10	2.04	2.01
	W2	2.12	2.14	2.13	2.19	2.09	2.37	2.20	2.43	2.11	2.13	2.22	2.22
CA1-	W3	2.53	2.53	2.58	2.63	2.58	2.59	2.52	2.57	2.51	2.58	2.61	2.56
	W3B/C										3.25	4.26	
	W4	2.36	2.30	2.37	2.22	2.57	2.37	2.37	2.43	2.30	2.30	2.20	2.29

*The O5–Ox distance in the 2F (400 μs) data set is shorter than in the 2F data set as for this data set we did not model two configurations for the OEC. Hence, the O5 position that is used to measure the O5–Ox distance is definitely influenced by the contribution of O5 in the S₂ state (closer to Mn1) and only partly by O5 in the S₃ state (longer Mn1–O5 distance).

Extended Data Table 3 | Channel nomenclature in the literature

This work	Ho and Styring ⁶⁶	Vassiliev ⁶⁷	Murray ⁶⁸	Gabdulkhakov ⁶⁹	Umena ⁷⁰	Sakashita ⁷¹
01 channel	large channel	4.A	<i>channel ii</i>	B1		01-water chain
04 channel	narrow channel	2		E, F		04-water chain
CI1 channel	broad channel	1	<i>channel iii</i>	D	4.b	E65/E312 channel
CI2 channel					4.c	

Summary of the correspondence of the multiple names used for identifying the water and proton channels in PSII⁶⁶⁻⁷¹.

Cryo-EM structures of a human ABCG2 mutant trapped in ATP-bound and substrate-bound states

Ioannis Manolaridis^{1,4}, Scott M. Jackson^{1,4}, Nicholas M. I. Taylor^{2,3,4}, Julia Kowal^{1,4}, Henning Stahlberg^{2*} & Kaspar P. Locher^{1*}

ABCG2 is a transporter protein of the ATP-binding-cassette (ABC) family that is expressed in the plasma membrane in cells of various tissues and tissue barriers, including the blood–brain, blood–testis and maternal–fetal barriers^{1–4}. Powered by ATP, it translocates endogenous substrates, affects the pharmacokinetics of many drugs and protects against a wide array of xenobiotics, including anti-cancer drugs^{5–12}. Previous studies have revealed the architecture of ABCG2 and the structural basis of its inhibition by small molecules and antibodies^{13,14}. However, the mechanisms of substrate recognition and ATP-driven transport are unknown. Here we present high-resolution cryo-electron microscopy (cryo-EM) structures of human ABCG2 in a substrate-bound pre-translocation state and an ATP-bound post-translocation state. For both structures, we used a mutant containing a glutamine replacing the catalytic glutamate (ABCG2_{EQ}), which resulted in reduced ATPase and transport rates and facilitated conformational trapping for structural studies. In the substrate-bound state, a single molecule of estrone-3-sulfate (E₁S) is bound in a central, hydrophobic and cytoplasm-facing cavity about halfway across the membrane. Only one molecule of E₁S can bind in the observed binding mode. In the ATP-bound state, the substrate-binding cavity has collapsed while an external cavity has opened to the extracellular side of the membrane. The ATP-induced conformational changes include rigid-body shifts of the transmembrane domains, pivoting of the nucleotide-binding domains (NBDs), and a change in the relative orientation of the NBD subdomains. Mutagenesis and in vitro characterization of transport and ATPase activities demonstrate the roles of specific residues in substrate recognition, including a leucine residue that forms a ‘plug’ between the two cavities. Our results show how ABCG2 harnesses the energy of ATP binding to extrude E₁S and other substrates, and suggest that the size and binding affinity of compounds are important for distinguishing substrates from inhibitors.

We first established that replacing the catalytic glutamate E211 with a glutamine in the Walker B motif (a phosphate-binding sequence) resulted in greatly reduced, but not abolished, ATP hydrolysis and E₁S transport activity¹³ (Fig. 1 and Extended Data Fig. 1). Next, to determine the E₁S-bound structure (ABCG2_{EQ}–E₁S), we added the antigen-binding fragment of the monoclonal antibody 5D3 (5D3-Fab) to the sample, which bound to the external side of ABCG2 and facilitated the determination of the high-resolution structure¹⁵. 5D3-Fab inhibits the transport activity of liposome-reconstituted ABCG2 and slows down its ATP hydrolysis, but has no effect on the half-maximal effective concentration (EC₅₀) of E₁S-induced ATPase stimulation, suggesting that it does not alter the interaction between ABCG2 and E₁S^{13,16} (Extended Data Fig. 2). The predominant three-dimensional (3D) class of nanodisc-reconstituted ABCG2_{EQ}–E₁S revealed an inward-open conformation and was refined to an overall resolution of 3.6 Å, such that the transmembrane domains (TMDs)—including the substrate-binding cavity—were clearly resolved (Extended Data Figs. 3, 4a and Extended Data Table 1). We observed a density feature in the substrate-binding cavity, which is formed by transmembrane (TM)

helices TM2 and TM5a of opposing ABCG2 monomers. The density could fit only one E₁S molecule but given that ABCG2 has two-fold symmetry, E₁S can be bound in two orientations related by a 180° rotation (Fig. 2a, b and Extended Data Fig. 4b). Two E₁S molecules cannot bind simultaneously because their polycyclic ring systems would clash sterically. The strongest density was at the two-fold symmetry axis, where the core of the flat polycyclic ring binds, and reprocessing the data with C1 symmetry resulted in a very similar—albeit lower-resolution—electron microscopy map (Extended Data Fig. 4b, c). The substrate-binding cavity was shown previously to accommodate potent inhibitors, demonstrating its dual role in substrate and multidrug binding¹⁴ (Fig. 2c).

The ABCG2_{EQ}–E₁S structure revealed which residues interacted with bound substrate (Fig. 2d). We generated single point mutations of all of these residues, and determined the in vitro ATPase and E₁S transport activities of the resulting ABCG2 variants upon reconstitution of the purified proteins in proteoliposomes (Fig. 2e, f). The stability of all mutants tested was similar to that of the wild-type protein (Extended Data Fig. 5a), allowing direct comparison. We also determined the EC₅₀ values of E₁S-induced ATPase stimulation for all mutants (Extended Data Fig. 5b, c). Consistent with their role in binding E₁S, the transport activities of the mutants N436A and F439A were strongly reduced, as were their ATPase activities. Notably, neither the N436A nor the F439A mutant showed stimulation of their ATPase activity by E₁S, indicating that the interactions suggested by the structure (a hydrogen bond between N436 and the sulfate group of E₁S, and the stacking interaction of the phenyl ring of F439 to the ring system of E₁S) are important for substrate binding (Fig. 2 and Extended Data Fig. 5b, c). The V546F mutant had impaired transport activity but displayed a 12-fold increase in basal ATPase activity, which was inhibited by E₁S in a concentration-dependent manner. This could suggest that the introduction of two phenyl rings at this position of the substrate-binding cavity mimics the binding of a substrate and thus stimulates ATPase activity, whereas further addition of E₁S ‘clogs’ the transporter.

The V546A mutant, by contrast, had similar functional characteristics to the wild-type protein, with a slight increase in the EC₅₀ of E₁S stimulation. It has previously been reported that in the ABCG5/ABCG8 heterodimeric protein, the mutations Y432A and A540F (equivalent to F439A and V546F in ABCG2) disrupted cholesterol transport¹⁷, suggesting a common location for the substrate-binding site among G-subfamily ABC transporters. We further found that the mutation T435A caused a roughly 4.5-fold increase in the apparent EC₅₀ of ATPase stimulation, consistent with our interpretation from the structure that a hydrogen bond exists between the β-hydroxyl group of T435 and the ester group of E₁S. There was a twofold increase in transport in the T435A mutant, suggesting an inverse relationship between binding affinity and the maximal transport rate. The introduction of two phenylalanines (T435F mutant) impaired both E₁S transport and ATP hydrolysis—a different effect to that seen with V546F. This emphasizes the sensitivity of the binding cavity to modifications. Finally, the M549A mutant had similar ATPase and transport activities

¹Institute of Molecular Biology and Biophysics, Department of Biology, ETH Zurich, Switzerland. ²Center for Cellular Imaging and NanoAnalytics (C-CINA), Biozentrum, University of Basel, Basel, Switzerland. ³Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. ⁴These authors contributed equally: Ioannis Manolaridis, Scott M. Jackson, Nicholas M. I. Taylor, Julia Kowal. *e-mail: henning.stahlberg@unibas.ch; locher@mol.biol.ethz.ch

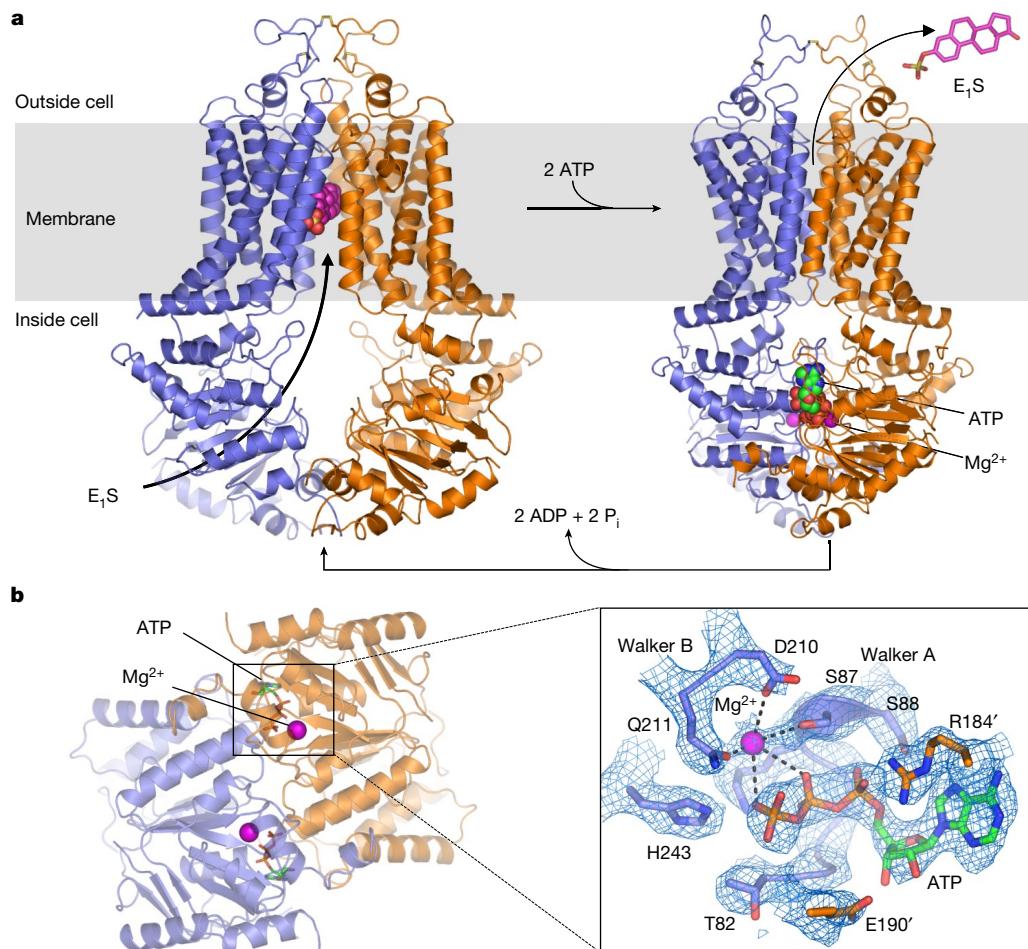


Fig. 1 | Structures and transport cycle of ABCG2. **a**, Cartoon representation of ABCG2_{EQ}–E₁S (left) and ABCG2_{EQ}–ATP (right). ABCG2 monomers coloured blue and orange. Bound E₁S, ATP and Mg²⁺ shown as spheres. In ABCG2_{EQ}–E₁S, bound 5D3-Fab omitted for clarity. **b**, Structure of NBD dimer from the ATP-bound state, viewed from

the cytoplasm, with bound ATP (sticks) and Mg²⁺ ions (spheres). Inset (rotated about 150° to the right, viewed from the membrane): electron microscopy density around bound ATP, with Walker A and Walker B motifs, E190 of the signature motif and the ‘switch’ histidine (H243) shown as sticks and labelled; Mg²⁺ shown as purple sphere.

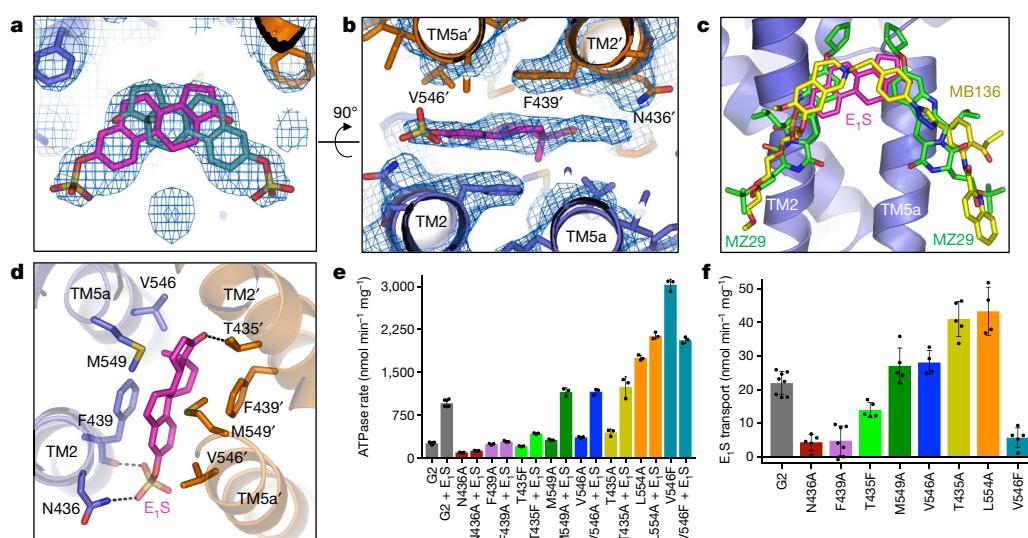


Fig. 2 | Substrate-binding cavity and mutant analysis. **a**, C2-symmetrized electron microscopy density of ABCG2_{EQ}–E₁S; bound E₁S molecule (pink or turquoise sticks) shown in two possible orientations, rotated by 180° along *y*-axis. **b**, As in **a**, but rotated 90° and showing one E₁S molecule and surrounding residues as viewed from the cytoplasm. TM helices and contacting residues are labelled. **c**, Overlay of E₁S (pink sticks) and inhibitors MZ29 (green sticks; Protein Data Bank accession number (PDB): 6ETI) and MB136 (yellow sticks; PDB: 6FEQ), bound in

the substrate-binding cavity, after superposition of the three structures. **d**, Substrate-binding cavity viewed from within the membrane, showing side chains (sticks) of residues interacting with E₁S (pink sticks). **e**, ATPase activities of liposome-reconstituted wild-type and mutant ABCG2 in the presence and absence of 50 μM E₁S. **f**, Initial E₁S-transport activities. The bars show means; error bars show standard deviations; and dots show rates derived from each technical replicate (same batch of liposomes).

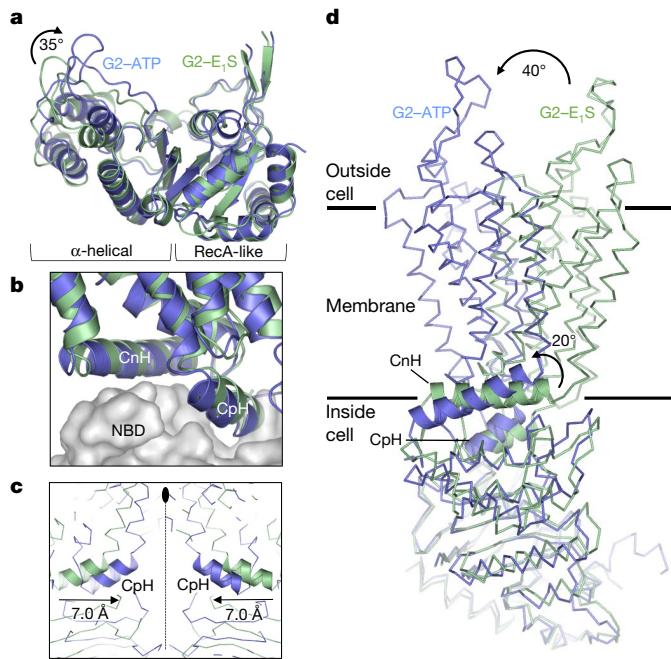


Fig. 3 | ATP-induced conformational changes. **a**, Superposition of the RecA-like subdomains of the NBDs of the ABCG2_{EQ}–E₁S (green) and ABCG2_{EQ}–ATP (blue) structures. A roughly 35° inward rotation of the helical subdomain is observed upon ATP binding. **b**, Superposition of one ABCG2 monomer of the ABCG2_{EQ}–E₁S and ABCG2_{EQ}–ATP structures, with the NBDs shown as a grey surface and the TMDs as ribbons. The interface helices—CpH and CnH—are labelled. **c**, Superposition of the ABCG2_{EQ}–E₁S and ABCG2_{EQ}–ATP structures along the two-fold symmetry axis (dotted line), showing a 7 Å inward movement of the CpH helices of each ABCG2 monomer. **d**, Comparison of a single ABCG2 monomer of the ABCG2_{EQ}–E₁S and ABCG2_{EQ}–ATP structures. The NBDs have been superimposed, revealing a 20° rotation of the CnH and CpH helices (shown as ribbons) as well as a 40° rotation of the TMDs relative to one another.

to the wild-type protein, suggesting a minor contribution of this residue to E₁S binding.

To visualize ATP-driven conformational changes in ABCG2, we added ATP and magnesium to nanodisc-reconstituted ABCG2_{EQ} in the absence of 5D3-Fab (ABCG2_{EQ}–ATP). Cryo-EM analysis revealed that most particles featured an ATP-bound conformation with the NBD dimer closed, and no inward-facing classes (Fig. 1 and Extended Data Fig. 6). The overall resolution was 3.1 Å, with excellent side-chain density for the TMDs, NBDs and nucleotides (Extended Data Fig. 4d and Extended Data Table 1). The structure revealed a closed, ‘head-to-tail’ NBD dimer, featuring a much larger interface than in the nucleotide-free state, and forming two ATP-binding sites between the Walker A motif (another phosphate-binding loop, or P-loop) of one NBD and the signature sequence (VSGGE sequence) of the other (Fig. 1b). ATP molecules are bound, and there is clear electron microscopy density for a magnesium ion interacting with the β- and γ-phosphates of each ATP (Fig. 1b). Three conserved side chains coordinate the γ-phosphate of ATP: Q211 (corresponding to the catalytic glutamate in wild-type ABCG2); H243 (corresponding to the ‘switch’ histidine¹⁸); and Q126 (which is part of the Q-loop). Q211 also coordinates the magnesium ion. ABCG2 does not contain an A-loop with an aromatic side chain stacking against the adenine moiety, as is seen in many other ABC transporters^{19–22}. Rather, one face of the adenine ring is in Van der Waals distance of residues V46, I63 and G185 from one NBD; the other face stacks against R184 from the opposite NBD. R184 also forms a salt bridge with the α-phosphate, which was observed previously in the AMPPNP-bound structure of the bacterial B12 transporter BtuCDF²³. The NBD interface of ATP-bound ABCG2_{EQ} also contains a salt bridge formed by E127 (part of the Q-loop) and R191 adjacent to the signature

motif. Unlike in B-family ABC transporters, there is a hole at the interface of the four domains of ABCG2, rather similar to what was seen in BtuCDF (Fig. 1a).

As a result of ATP binding, the α-helical domains of the NBDs have rotated by roughly 35° relative to the RecA-like domains, approaching the opposite NBD and the two-fold symmetry axis (Fig. 3a). This rotation is required for NBD dimerization and is part of the ‘power stroke’ in the transport cycle²⁴. The individual TMD–NBD interfaces—formed mainly by the ‘connecting helices’ (CnH or TM1a) and ‘coupling helices’ (CpH, corresponding to the carboxy-terminal part of TM2)—remain largely unchanged in each ABCG2 monomer between the nucleotide-free and ATP-bound states (Fig. 3b). However, because of the shift in the NBDs, the cytoplasmic parts of the TMDs are pushed towards each other, with each CpH approaching the two-fold symmetry axis by around 7 Å (Fig. 3c). The altered conformations of the TMDs can be described as rigid-body movements, with CnH and CpH acting as the pivot points for the transition. These two α-helices undergo a roughly 20° rotation when superposing the NBDs in the structures of the two states, which translates to a roughly 40° rotation of the helical axes of the TMDs (Fig. 3d).

The ATP-induced conformational changes have important consequences for the substrate-translocation pathway. In the ABCG2_{EQ}–E₁S structure, the phenyl rings of the F439 residues of the two ABCG2 monomers are 8.0 Å apart, with bound E₁S between them (Fig. 2d). By contrast, these phenyl rings stack against each other in the ATP-bound state (Fig. 4a, c) and the substrate-binding cavity has completely collapsed, with no space for bound substrate. To be transported across the membrane, therefore, the substrate has to move through the likely translocation pathway at the centre of the transporter and reach the external cavity before the pathway is completely closed. This can be accomplished only if there are transient conformational changes such as TM-helix bending, to generate space for the substrate. Such transient changes resemble a peristaltic motion.

The external cavity—occluded in the nucleotide-free state—is open to the outside in the ATP-bound state, while maintaining the intra- and intermolecular disulfides (C592–C608 and C603–C603') in extracellular loop 3 (EL3), promoting substrate release (Figs. 1a, 4d and Extended Data Figs. 1b, 6e). Two leucine residues (L554 and L555), in the loop between TM5a and TM5b, form a plug that separates the substrate-binding cavity and the external cavity (Fig. 4). We individually mutated these leucines to alanines and found that the L555A variant did not express any functional protein, suggesting a structural role of L555 in addition to a likely gating function. By contrast, the L554A mutant was stable and showed functional differences compared to wild-type ABCG2 (Fig. 2e, f and Extended Data Fig. 5): the basal ATPase rate of L554A was greatly increased and there was only a minor stimulation of the ATPase rate by E₁S (around 20%, compared with roughly 3.5-fold in wild-type ABCG2). Furthermore, the apparent EC₅₀ of E₁S-induced ATPase stimulation was increased, suggesting weaker substrate binding. Finally, the E₁S transport activity of the L554A mutant was twice as high as that of wild-type ABCG2. We interpret that the opening and closing of the ‘leucine plug’ may act as a checkpoint during the transport reaction, and that although the removal of the leucine side chain accelerates the transport process, it may reduce substrate selectivity.

A comparison of the structures shown here provides insight into the transport cycle of ABCG2. Substrate may bind via the cytoplasm or from within the lipid bilayer via the ‘membrane entrance’¹⁴ (Fig. 4b). Once substrate is bound, the NBD dimer can only close when the substrate moves out of the substrate-binding cavity, because this cavity does not provide any space when ATP is also bound. In a productive transport cycle, substrate probably moves through a translocation pathway at the centre of the transporter, via the ‘leucine plug’. The structure of ATP-bound ABCG2_{EQ} suggests that once a substrate clears the plug area and enters the external cavity, the plug region closes and substrate is released to the outside (Fig. 4d). One caveat is that the E211Q mutation may have influenced the energetics of the conformational changes

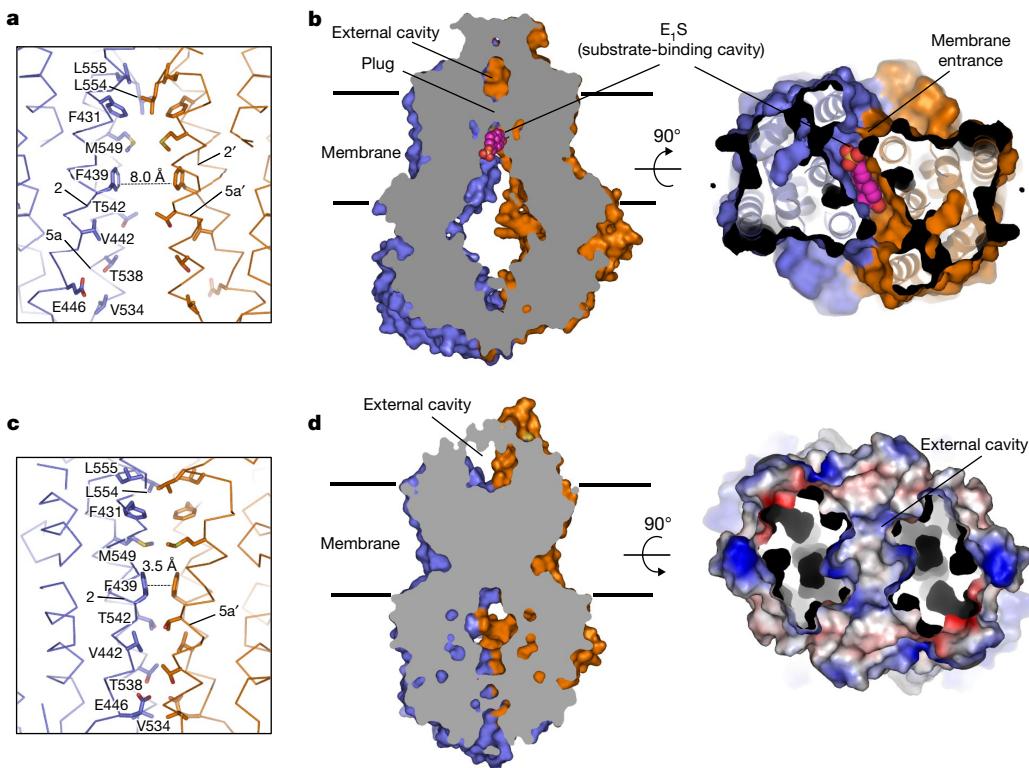


Fig. 4 | Substrate-translocation pathway. **a**, $\text{C}\alpha$ trace of the translocation-pathway region of $\text{ABCG2}_{\text{EQ}}\text{--E}_1\text{S}$. Residues lining the substrate-binding cavity are shown as sticks; bound E_1S has been omitted for clarity. The dashed line shows the distance between the two F439 residues that stack against bound E_1S . **b**, Vertical slice through a surface representation of $\text{ABCG2}_{\text{EQ}}\text{--E}_1\text{S}$, with bound E_1S shown as pink spheres and the two cavities and plug region labelled. In the right-hand panel, a 90° rotation of the

structure reveals the fit of E_1S in the substrate-binding cavity, as viewed from the cytoplasm. The NBDs have been removed for clarity. **c, d**, As for **a, b**, but with the $\text{ABCG2}_{\text{EQ}}\text{--ATP}$ structure. In the right-hand panel of **d**, the molecular surface of the external cavity, viewed from the extracellular space and colour-coded by electrostatic potential ranging from blue (most positive) to red (most negative), is shown with the extracellular loop EL3 removed for clarity.

involved. Our findings suggest that ATP binding might be sufficient for the substrate-extrusion step and that ATP hydrolysis might be required to reset the transporter to an inward-facing conformation^{19,20,25,26}. Unlike many other transporters^{22,27–30}, ABCG2 does not appear to form a stable, occluded conformation providing space for bound substrate, but rather a transient conformation that is consistent with a peristalsis-like mechanism, reminiscent of the bacterial BtuCDF transporter.

Given their polyspecificity, a key unanswered question concerning multidrug transporters such as ABCG2 is why certain compounds act as substrates, while others are potent inhibitors. Our results allow us to compare the binding modes of a bona fide ABCG2 substrate with those of two potent inhibitors. All three molecules bind in the same cavity of the transporter (Fig. 2c). However, whereas a single E_1S molecule binds on the two-fold symmetry axis and deep in the cavity, two molecules of MZ29—a compound derived from the ABCG2 inhibitor Ko143—have been found to fill the substrate-binding cavity completely, almost reaching the cytoplasmic membrane boundary and forming many additional contacts with the surface of ABCG2 (ref. ¹⁴). A single copy of the tariquidar-derived inhibitor MB136 also fills the binding cavity completely, forming similar contacts to MZ29. The numerous additional contacts, which include residues of TM1b in addition to TM2 and TM5a, can explain the increased binding affinity of inhibitors compared with substrates. The difference appears to mean that when E_1S and ATP are bound to ABCG2, the NBDs and the cytoplasmic part of the TMDs can still approach, allowing for an opening of the plug and simultaneous pushing of the substrate into the external cavity. Inhibitors, by contrast, act as ‘wedges’ and immobilize the transporter by locking it in an inward-facing conformation. The reduced size, binding surface and affinity of substrates, and their binding deeper inside the substrate cavity, allow them to be translocated in a productive transport cycle.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0680-3>.

Received: 26 March 2018; Accepted: 24 August 2018;
Published online 7 November 2018.

1. Diestra, J. E. et al. Frequent expression of the multi-drug resistance-associated protein BCRP/MXR/ABCP/ABCG2 in human tumours detected by the BXP-21 monoclonal antibody in paraffin-embedded material. *J. Pathol.* **198**, 213–219 (2002).
2. Fetsch, P. A. et al. Localization of the ABCG2 mitoxantrone resistance-associated protein in normal tissues. *Cancer Lett.* **235**, 84–92 (2006).
3. Robey, R. W. et al. ABCG2: a perspective. *Adv. Drug Deliv. Rev.* **61**, 3–13 (2009).
4. Vlaming, M. L., Lagas, J. S. & Schinkel, A. H. Physiological and pharmacological roles of ABCG2 (BCRP): recent findings in Abcg2 knockout mice. *Adv. Drug Deliv. Rev.* **61**, 14–25 (2009).
5. Gillet, J. P. & Gottesman, M. M. Advances in the molecular detection of ABC transporters involved in multidrug resistance in cancer. *Curr. Pharm. Biotechnol.* **12**, 686–692 (2011).
6. Gottesman, M. M., Fojo, T. & Bates, S. E. Multidrug resistance in cancer: role of ATP-dependent transporters. *Nat. Rev. Cancer* **2**, 48–58 (2002).
7. Imai, Y. et al. Breast cancer resistance protein exports sulfated estrogens but not free estrogens. *Mol. Pharmacol.* **64**, 610–618 (2003).
8. Ishikawa, T., Aw, W. & Kaneko, K. Metabolic interactions of purine derivatives with human ABC transporter ABCG2: genetic testing to assess gout risk. *Pharmaceuticals* **6**, 1347–1360 (2013).
9. Mao, Q. & Unadkat, J. D. Role of the breast cancer resistance protein (BCRP/ABCG2) in drug transport—an update. *AAPS J.* **17**, 65–82 (2015).
10. Mo, W. & Zhang, J. T. Human ABCG2: structure, function, and its role in multidrug resistance. *Int. J. Biochem. Mol. Biol.* **3**, 1–27 (2012).
11. Sarkadi, B., Homolya, L., Szakacs, G. & Váradai, A. Human multidrug resistance ABCB and ABCG transporters: participation in a chemoimmunity defense system. *Physiol. Rev.* **86**, 1179–1236 (2006).

12. Sharom, F. J. The P-glycoprotein multidrug transporter. *Essays Biochem.* **50**, 161–178 (2011).
13. Taylor, N. M. I. et al. Structure of the human multidrug transporter ABCG2. *Nature* **546**, 504–509 (2017).
14. Jackson, S. M. et al. Structural basis of small-molecule inhibition of human multidrug transporter ABCG2. *Nat. Struct. Mol. Biol.* **25**, 333–340 (2018).
15. Zhou, S. et al. The ABC transporter Bcrp1/ABCG2 is expressed in a wide variety of stem cells and is a molecular determinant of the side-population phenotype. *Nat. Med.* **7**, 1028–1034 (2001).
16. Suzuki, M., Suzuki, H., Sugimoto, Y. & Sugiyama, Y. ABCG2 transports sulfated conjugates of steroids and xenobiotics. *J. Biol. Chem.* **278**, 22644–22649 (2003).
17. Lee, J. Y. et al. Crystal structure of the human sterol transporter ABCG5/ABCG8. *Nature* **533**, 561–564 (2016).
18. Hanekop, N., Zaitseva, J., Jenewein, S., Holland, I. B. & Schmitt, L. Molecular insights into the mechanism of ATP-hydrolysis by the NBD of the ABC-transporter HlyB. *FEBS Lett.* **580**, 1036–1041 (2006).
19. Johnson, Z. L. & Chen, J. ATP binding enables substrate release from multidrug resistance protein 1. *Cell* **172**, 81–89 (2018).
20. Kim, Y. & Chen, J. Molecular structure of human P-glycoprotein in the ATP-bound, outward-facing conformation. *Science* **359**, 915–919 (2018).
21. Shintre, C. A. et al. Structures of ABCB10, a human ATP-binding cassette transporter in apo- and nucleotide-bound states. *Proc. Natl. Acad. Sci. USA* **110**, 9710–9715 (2013).
22. Zhang, Z., Liu, F. & Chen, J. Conformational changes of CFTR upon phosphorylation and ATP binding. *Cell* **170**, 483–491 (2017).
23. Korkhov, V. M., Mireku, S. A. & Locher, K. P. Structure of AMP-PNP-bound vitamin B12 transporter BtuCD-F. *Nature* **490**, 367–372 (2012).
24. Zaitseva, J. et al. A structural analysis of asymmetry required for catalytic activity of an ABC-ATPase domain dimer. *EMBO J.* **25**, 3432–3443 (2006).
25. Aller, S. G. et al. Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding. *Science* **323**, 1718–1722 (2009).
26. Johnson, Z. L. & Chen, J. Structural basis of substrate recognition by the multidrug resistance protein MRP1. *Cell* **168**, 1075–1085 (2017).
27. Alam, A. et al. Structure of a zosuquidar and UIC2-bound human-mouse chimeric ABCB1. *Proc. Natl. Acad. Sci. USA* **115**, E1973–E1982 (2018).
28. Choudhury, H. G. et al. Structure of an antibacterial peptide ATP-binding cassette transporter in a novel outward occluded state. *Proc. Natl. Acad. Sci. USA* **111**, 9145–9150 (2014).
29. Hohl, M., Briand, C., Grüter, M. G. & Seeger, M. A. Crystal structure of a heterodimeric ABC transporter in its inward-facing conformation. *Nat. Struct. Mol. Biol.* **19**, 395–402 (2012).
30. Verhagen, B. et al. Energy transduction and alternating access of the mammalian ABC transporter P-glycoprotein. *Nature* **543**, 738–741 (2017).

Acknowledgements This research was supported by the Swiss National Science Foundation through the National Centre of Competence in Research (NCCR) TransCure and by a Swiss Federal Institute of Technology Zurich (ETH Zurich) research grant (ETH-22-14-1). N.M.I.T. was also supported by the Research Fund Junior Researchers of the University of Basel. J.K. was also supported by the TransCure Young Investigator Award (2017). Cryo-EM data were collected at C-CINA, University of Basel; we thank K. Goldie, L. Kováčik and A. Fecteau-Lefebvre for technical support. We thank N. Tremp for help with cell culture and B. Sorrentino (St Jude Children's Research Hospital) for providing the 5D3-producing hybridoma cell line.

Reviewer information *Nature* thanks H. Mchaourab and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions I.M. expressed and purified wild-type ABCG2 and 5D3-Fab. I.M. and S.M.J. cloned, expressed and purified the ABCG2 mutants. S.M.J. reconstituted ABCG2 into liposomes and lipidic nanodiscs for cryo-EM and functional studies and carried out all functional experiments. J.K. prepared cryo-grids. N.M.I.T. collected cryo-EM data with the assistance of H.S. I.M. processed cryo-EM data of ATP-bound ABCG2 and determined the structure with the assistance of J.K. N.M.I.T. processed electron microscopy data and determined the structure of E₁S-bound ABCG2. I.M and K.P.L. built, refined and validated the structures. K.P.L., I.M. and S.M.J. conceived the project, designed the experiments and wrote the manuscript. All authors contributed to revision of the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0680-3>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0680-3>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to H.S. or K.P.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Expression and purification of wild-type and mutant human ABCG2. Human ABCG2, containing an amino (N)-terminal Flag tag, was expressed in HEK293-EBNA (Thermo Fisher Scientific) cells and purified as described^{13,14}.

Expression and purification of 5D3-Fab. 5D3 hybridoma cells, producing the 5D3 monoclonal antibody, were obtained from B. Sorrentino. The cells were cultured in WHEATON CELLine Bioreactors, according to the manufacturer's protocol, and 5D3-Fab was then purified from the supernatant, as described in the Fab Preparation Kit protocol (Thermo Fisher Scientific).

Nanodisc preparation of ABCG2_{EQ}. The membrane scaffold protein (MSP) 1D1 was expressed and purified³¹ and ABCG2 was reconstituted into brain polar lipid (BPL)/cholesterol hemisuccinate (CHS) nanodiscs as described^{13,14}. To generate the ABCG2_{EQ}-E₁S sample for cryo-EM studies, ABCG2 was first mixed with a threefold molar excess of 5D3-Fab before reconstitution. After size-exclusion chromatography (SEC) using a Superdex 200 10/300 column (GE Healthcare), the complex was incubated with 200 μM E₁S, 5 mM ATP and 5 mM MgCl₂ for 15 min at room temperature before grid preparation. For the ABCG2_{EQ}-ATP sample, following SEC, the complex was incubated with 5 mM ATP and 5 mM MgCl₂ for 15 min at room temperature before grid preparation.

ABCG2 liposome preparation. A BPL/cholesterol (Avanti Polar Lipids) mixture was prepared at a 4/1 (w/w) ratio as described³². Detergent-purified ABCG2 was then mixed with BPL/cholesterol; detergent was removed with Bio-Beads and the reconstitution efficiency determined^{13,14,33}.

Transport assays. In vitro transport assays using ABCG2 proteoliposomes, containing either wild-type or mutant protein, were carried out as described. In brief, ABCG2 proteoliposomes were extruded and then incubated with 5 mM MgCl₂ and 50 μM [³H]-E₁S for 5 min at 30 °C. The transport reaction was initiated by the addition of 2 mM ATP and the sample was filtered using a Multiscreen vacuum manifold (MSFBN6B filter plate, Millipore). Radioactivity trapped on the filters was measured using a scintillation counter and the initial transport rates (over 30 seconds to 2 minutes) were determined using linear regression in GraphPad Prism 7.00. Rates were corrected for the orientation of ABCG2 in proteoliposomes^{13,14}.

ATPase assays and determination of the EC₅₀ of E₁S stimulation. ATP-hydrolysis activity was measured using a previously described technique³⁴. All reactions were performed at 37 °C in the presence of 2 mM ATP and 10 mM MgCl₂^{13,14}. For ATPase assays in proteoliposomes, experiments were completed in the presence of 0–300 μM E₁S. To assess the effect of 5D3-Fab, ABCG2 proteoliposomes were freeze-thawed five times in the presence of a threefold molar excess of 5D3-Fab (to ensure that the antibody was present inside the proteoliposomes) before extrusion. Assays in nanodiscs were performed in the absence of E₁S. Data were recorded at four time intervals (0, 5, 15 and 30 min) and subsequent ATPase rates were determined using linear regression in GraphPad Prism 7.00. Rates were corrected for the orientation of ABCG2 in proteoliposomes. To determine the EC₅₀ of E₁S stimulation, we plotted the ATPase rates against the E₁S concentration, and generated curves using the nonlinear regression Michaelis–Menten analysis tool in GraphPad Prism 7.00.

Sample preparation and cryo-EM data acquisition. All cryo grids of ABCG2_{EQ} were prepared using a Vitrobot Mark IV (FEI) with the environmental chamber set at 100% humidity and 4 °C. An aliquot of 4 μl purified ABCG2_{EQ}-E₁S or ABCG2_{EQ}-ATP, at a protein concentration of approximately 0.4 mg ml⁻¹, was applied to glow-discharged Quantifoil (1.2/1.3) 300-mesh copper grids. After being blotted with filter paper for 2.0 s, the grids were flash-frozen in a mixture of propane and ethane cooled with liquid nitrogen. The ABCG2_{EQ}-E₁S dataset was composed of 3,984 movies and the ABCG2_{EQ}-ATP dataset was composed of 4,905 movies. Cryo-EM image data were recorded using SerialEM³⁵ on a Titan Krios microscope, operated at 300 kV and equipped with a Gatan Quantum-LS energy filter (20 eV zero loss filtering), containing a K2 Summit direct electron detector. Images were recorded in super-resolution counting mode with a pixel size of 0.4058 Å per pixel. Exposures were 10 s, dose-fractionated into 50 frames (0.2 s per frame), resulting in a frame dose of 2.0 electrons per Å². Data-collection quality was monitored using Focus³⁶. The first frame of each movie was discarded. Stacks were gain-normalized, motion-corrected, dose-weighted and averaged and then Fourier-cropped twofold with MotionCor2 (ref. ³⁷). Defocus estimates were obtained on the non-dose-weighted micrographs with CTFFIND4 (ref. ³⁸) for ABCG2_{EQ}-E₁S and Gctf³⁹ for ABCG2_{EQ}-ATP.

Image processing. Particles were picked automatically using Gautomatch (<http://www.mrc-lmb.cam.ac.uk/kzhang/>), resulting in an ABCG2_{EQ}-E₁S dataset of 168,184 particles—processed with CryoSPARC⁴⁰—and an ABCG2_{EQ}-ATP dataset of 1,128,170 particles, processed with RELION⁴¹. In both cases, 2D classification yielded at least 12 representative classes, containing 62,616 particles for ABCG2_{EQ}-E₁S and 543,142 particles for ABCG2_{EQ}-ATP. These classes were used for ab initio reconstruction (applying either C1 or C2 symmetry).

For ABCG2_{EQ}-E₁S, the resulting 3D models were used as a starting point for heterogeneous 3D refinement with three classes. Subsequent homogeneous refinements of all three classes (applying C2 symmetry) separately yielded maps that were very alike, with the maps for the two largest classes (corresponding to 38.7% and 32.6% of particles) having very similar densities in the substrate-binding cavity of ABCG2, and the map of smaller class (corresponding to 28.5% of the particles) having some additional density underneath the substrate density (which was only visible at noise level in the other classes). Therefore, we combined particles of the two largest classes (42,790 particles in total) from the 3D heterogeneous refinement and refined (applying C2 symmetry) against the heterogeneous refinement map of the largest class. This resulted in a map with a resolution of 3.58 Å, which was sharpened with an automatically calculated B-factor of −82.6 Å². For validation, all refinements were also performed using C1 symmetry.

For ABCG2_{EQ}-ATP, initial 3D classification into three classes resulted in one outstanding class, containing 51.8% of particles (corresponding to 288,447 particles) with clear secondary-structure elements. This class was 3D-refined, applying C2 symmetry and a soft mask, resulting in a 3D reconstruction with an overall resolution of 3.14 Å. For nanodisc subtraction, the ABCG2_{EQ}-ATP map was segmented in Chimera (using Segger) and the resulting nanodisc-only map was used to calculate projections, which were subsequently subtracted from the experimental particles. A soft mask, in which the subtracted density from the experimental particles was white (1) and the rest of the protein and the solvent were black (0), was generated by low-passing the nanodisc map to 15 Å and expanding the mask by 4 pixels with a soft edge of 6 pixels. The nanodisc-free experimental particles were used in further 3D refinements, which resulted in an electron microscopy map at a resolution of 3.09 Å (B-factor of −136 Å²), using the low-passed ABCG2_{EQ}-ATP map without nanodisc density as a reference. This map was used for ABCG2_{EQ}-ATP model building. Subsequent 3D classification in C1 symmetry, without applied alignments, was carried out in an effort to improve the EL3 density; however, despite extensive efforts the inherent flexibility of EL3 in this conformation prohibited its accurate modelling.

All resolutions were estimated with the Fourier shell correlation (FSC) 0.143 cut-off criterion⁴². ResMap⁴³ was used to calculate the local resolution maps.

Model building and refinement. For the ABCG2_{EQ}-E₁S structure, we used the post-processed map at an overall resolution of 3.58 Å. To generate an initial model, we docked the ABCG2-MZ29-Fab structure (Protein Data Bank 6ETI) into the electron microscopy density using Coot⁴⁴, and carried out manual fitting and modifications where the resolution allowed. The E₁S coordinates and restraints were generated using eLBOW⁴⁵ and fitted into the electron microscopy density using Coot. For the ABCG2_{EQ}-ATP structure, we used the post-processed, nanodisc-subtracted map at an overall resolution of 3.09 Å. The electron microscopy density was of excellent quality and allowed for the accurate building of ABCG2, ATP and magnesium from the ligand library in Coot, using the ABCG2-MZ29-Fab structure (PDB: 6ETI) as a template. For both structures, the complete models were refined against the working maps in PHENIX⁴⁶ using the program phenix.real_space_refine. For the final round of model refinement, we performed global real-space refinement with standard geometry restraints as well as rotamer, Ramachandran plot, C-beta, non-crystallographic symmetry and secondary-structure restraints, coupled to reciprocal-space refinement of the B-factors. The quality of the final models was analysed by MolProbity⁴⁷ and the refinement statistics are given in Extended Data Table 1. To validate the refinement, we introduced random shifts (mean value of 0.3 Å) into the coordinates of the final refined models using the program phenix.pdbtools⁴⁶, followed by refinement with phenix.real_space_refine (using the same parameters as described before) against the first unfiltered half-map (half-map 1). The overlay between the FSC curve of the model with random displacements refined against half-map 1 versus half-map 1 and the FSC curve of the same model versus half-map 2 (against which it was not refined) indicated that no over-refinement took place.

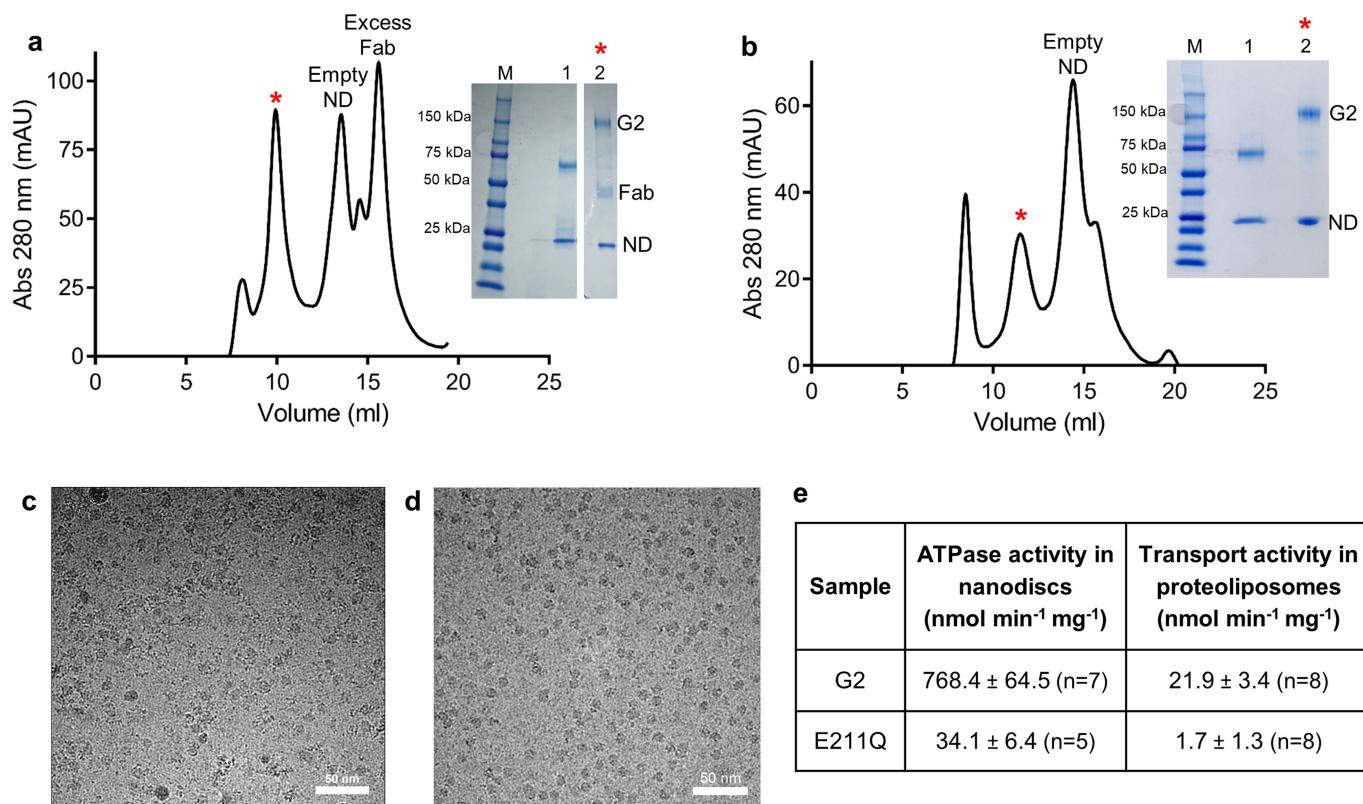
Figure preparation. Figures were prepared using the programs PyMOL (PyMOL Molecular Graphics System, DeLano Scientific) and GraphPad Prism 7.00 (GraphPad Software).

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

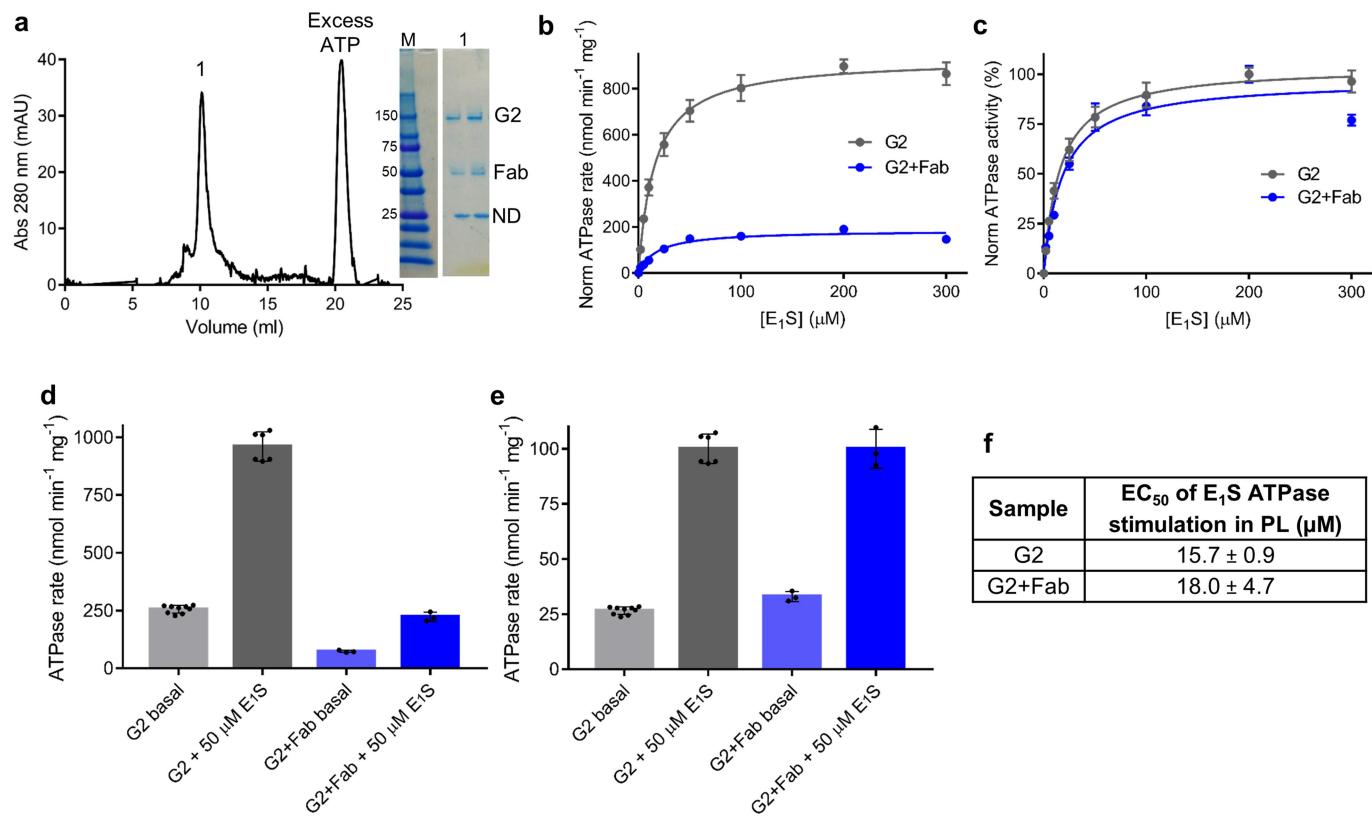
Atomic coordinates for ABCG2_{EQ}-E₁S (including only the variable domain of 5D3-Fab) and ABCG2_{EQ}-ATP were deposited in the Protein Data Bank under accession codes 6HCO and 6HBU, respectively. Electron microscopy data for the two structures were deposited in the Electron Microscopy Data Bank under accession codes EMD-0196 (ABCG2_{EQ}-E₁S) and EMD-0190 (ABCG2_{EQ}-ATP). Source Data for Fig. 2e, f and Extended Data Figs. 1e, 2b, d, f and 5 are available online. All other data are available from the corresponding author upon reasonable request. A Life Sciences Reporting Summary for this article is available.

31. Ritchie, T. K. et al. Chapter 11—reconstitution of membrane proteins in phospholipid bilayer nanodiscs. *Methods Enzymol.* **464**, 211–231 (2009).
32. Geertsma, E. R., Nik Mahmood, N. A., Schuurman-Wolters, G. K. & Poolman, B. Membrane reconstitution of ABC transporters and assays of translocator function. *Nat. Protocols* **3**, 256–266 (2008).
33. Schaffner, W. & Weissmann, C. A rapid, sensitive, and specific method for the determination of protein in dilute solution. *Anal. Biochem.* **56**, 502–514 (1973).
34. Chifflet, S., Torriglia, A., Chiesa, R. & Tolosa, S. A method for the determination of inorganic phosphate in the presence of labile organic phosphate and high concentrations of protein: application to lens ATPases. *Anal. Biochem.* **168**, 1–4 (1988).
35. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
36. Biyani, N. et al. Focus: the interface between data collection and data processing in cryo-EM. *J. Struct. Biol.* **198**, 124–133 (2017).
37. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
38. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
39. Zhang, K. Gctf: real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
40. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
41. Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5**, e18722 (2016).
42. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).
43. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
44. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
45. Moriarty, N. W., Grosse-Kunstleve, R. W. & Adams, P. D. Electronic ligand builder and optimization workbench (eLBOW): a tool for ligand coordinate and restraint generation. *Acta Crystallogr. D* **65**, 1074–1080 (2009).
46. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
47. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).



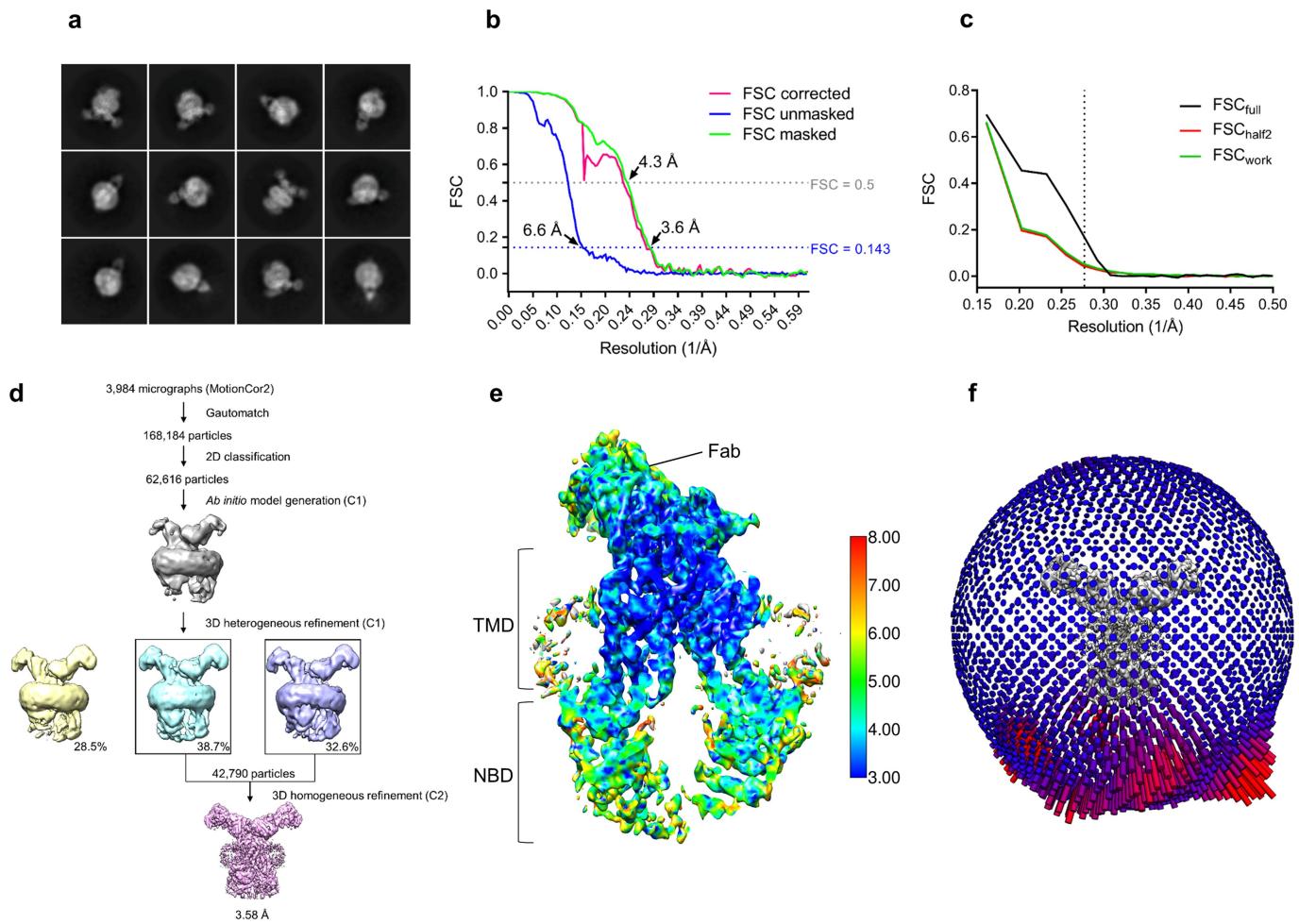
Extended Data Fig. 1 | Purification, activity and cryo-EM micrographs of ABCG2. **a**, Preparative SEC profile (milli absorbance units (mAU) at 280 nm plotted against retention volume (ml)) of the nanodisc-reconstituted ABCG2_{EQ}-E₁S complex. The fraction used for cryo-EM grid preparation is indicated by a red asterisk. Inset: reducing (lane 1) and non-reducing (lane 2) SDS-PAGE of the complex, showing bands for ABCG2 (G2), 5D3-Fab (Fab) and nanodisc (ND). **b**, Preparative SEC profile of the nanodisc-reconstituted ABCG2_{EQ}-ATP complex. The fraction used for cryo-EM grid preparation is indicated by a red asterisk. Inset: reducing (lane 1) and non-reducing (lane 2) SDS-PAGE of the complex, showing

bands for ABCG2 (G2) and nanodisc (ND). **c**, An example micrograph (drift-corrected, dose-weighted and low-pass-filtered to 20 Å) of the nanodisc-reconstituted ABCG2_{EQ}-E₁S sample. White scale bar, 50 nm. **d**, An example micrograph (drift-corrected, dose-weighted and low-pass-filtered to 20 Å) of the nanodisc-reconstituted ABCG2_{EQ}-ATP sample. White scale bar, 50 nm. **e**, ATPase activities of nanodisc-reconstituted and E₁S-transport activities of liposome-reconstituted ABCG2. In both cases, data for wild-type and mutant (E211Q) ABCG2 are shown. The standard deviation from *n* technical replicates (same batch of nanodiscs or liposomes) is shown.


Extended Data Fig. 2 | Effect of 5D3-Fab on ABCG2 function.

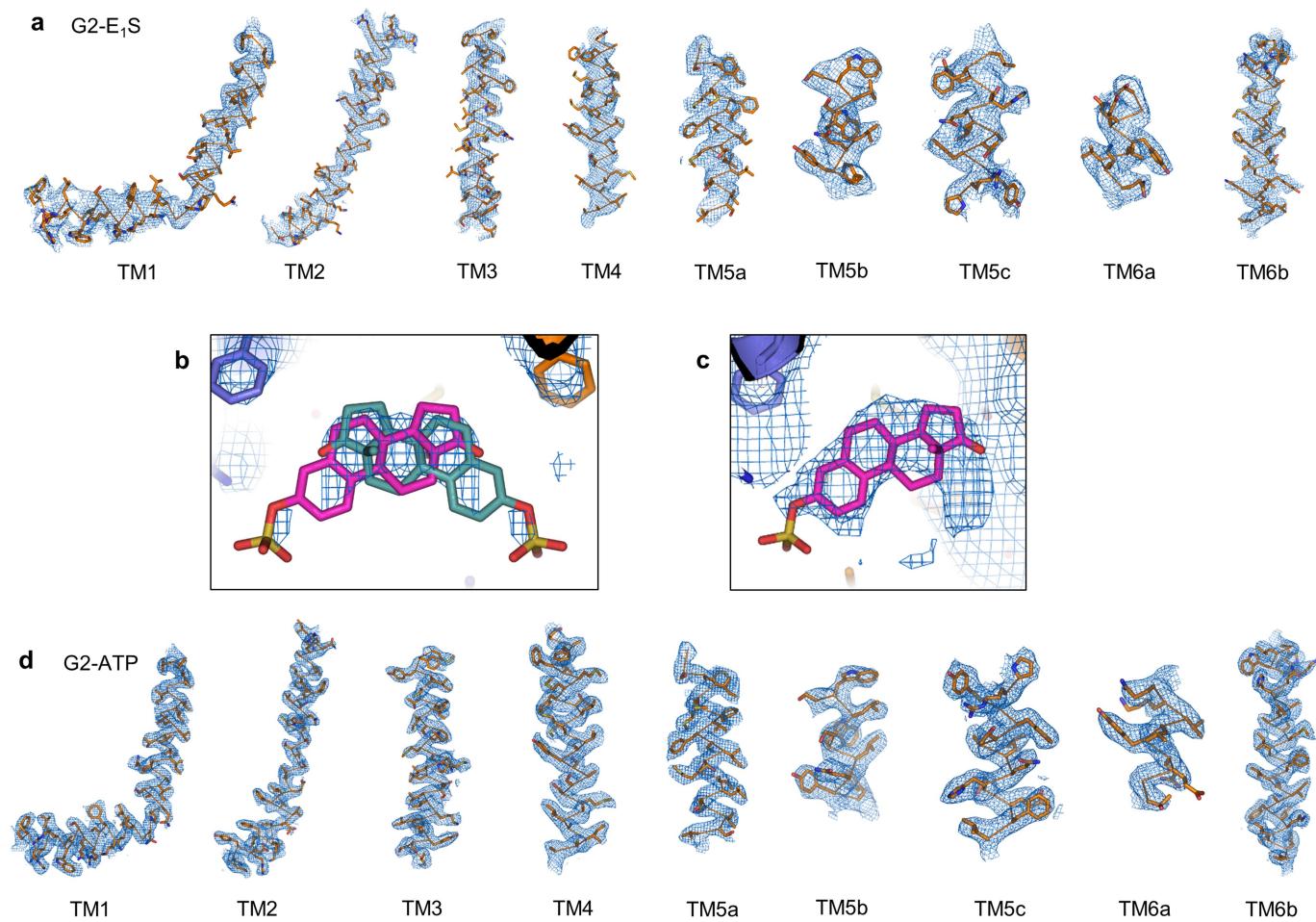
a, Analytical SEC profile of the nanodisc-reconstituted ABCG2_{EQ}–E₁S complex in the presence of 5 mM ATP and 5 mM MgCl₂. ‘1’ denotes the peak collected. Inset: non-reducing SDS–PAGE of the complex, showing bands for ABCG2 (G2), 5D3-Fab (Fab) and nanodisc (ND). **b**, ATPase activity of liposome-reconstituted ABCG2, in the presence or absence of 5D3-Fab, and with 0–300 μM E₁S. The basal ATPase activity has been normalized (norm) to 0. **c**, As for **b**, but with the maximal ATPase activity set to 100%. Each point represents the mean rate derived from technical

replicates. For G2 $n = 6$, except in the case of 0 and 200 μM E₁S, for which $n = 9$. For G2 + Fab, $n = 3$. **d**, ATPase activities of ABCG2 in the presence and absence of 5D3-Fab, and either 0 or 50 μM E₁S. **e**, As for **d**, but with activities in the presence of E₁S set to 100%. Bars show means and dots show the rates derived from each technical replicate (same batch of liposomes). Error bars show the standard deviation. **f**, The EC₅₀ of E₁S ATPase stimulation determined using the curves in **b** and **c** with the error of the fit (standard deviation) shown. PL, proteoliposome.



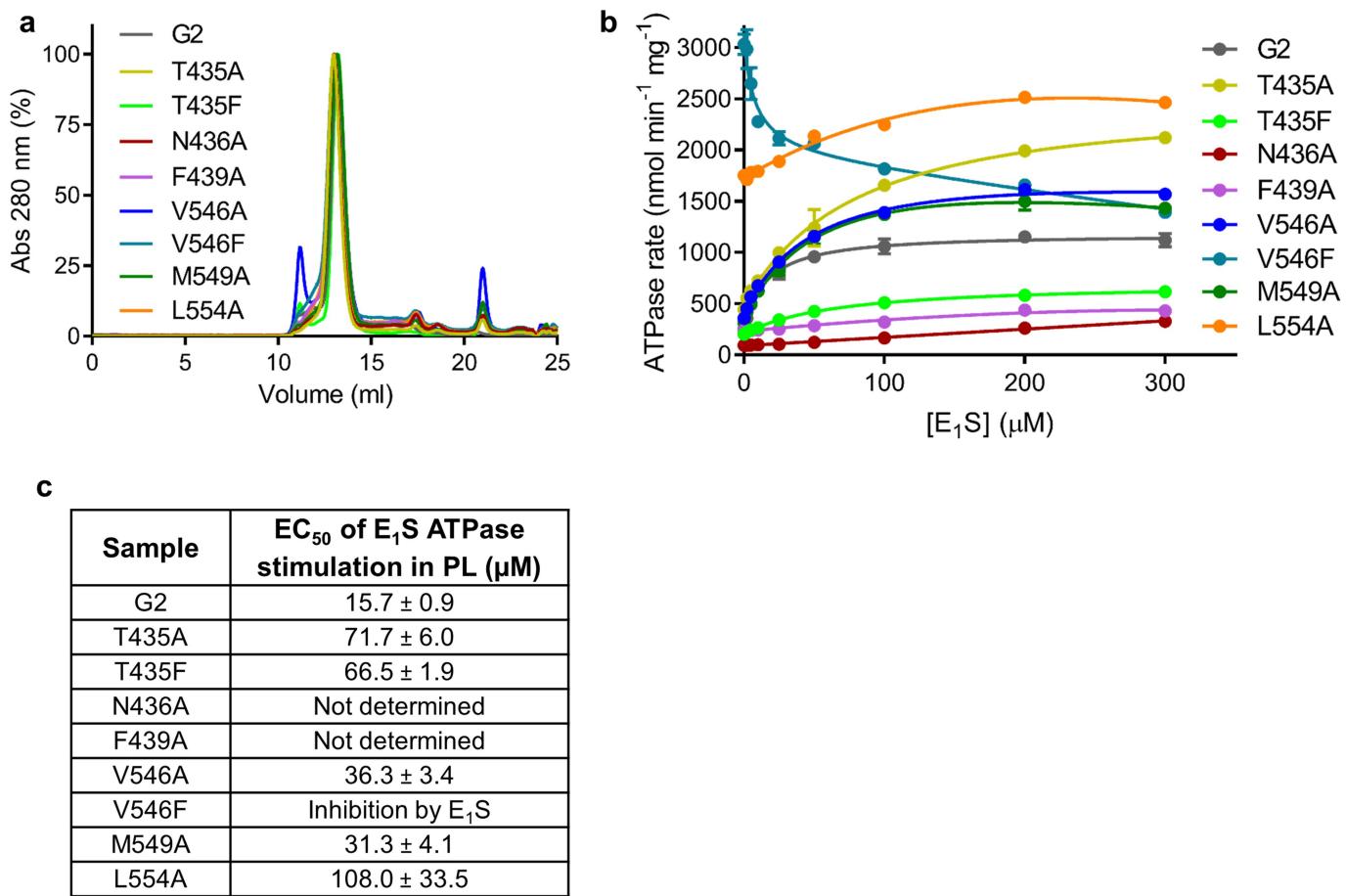
Extended Data Fig. 3 | Cryo-EM map generation, data processing and atomic-model refinement of ABCG2_{EQ}–E₁S. **a**, Twelve representative 2D class averages of the final round of 2D classification, sorted in decreasing order by the number of particles assigned to each class. **b**, FSC from the CryoSPARC auto-refine procedure of the unmasked half-maps (blue), the half-maps after masking (green), and the half-maps after masking and correction for the influence of the mask (pink). A horizontal dotted line (blue) is drawn for the FSC = 0.143 criterion. For both the unmasked and the corrected FSC curves, their intersection with the FSC = 0.143 and the FSC = 0.5 lines are marked by arrows, and the resolutions at these points are indicated. **c**, FSC curve of the final 3.58 Å refined model

versus the map against it was refined (FSC_{full}; black line). The FSC curve of the final refined model with introduced shifts (mean value of 0.3 Å) versus the first of two independent half-maps (half-map 1, against which it was refined; FSC_{work}; green line) or the same model versus the second independent half-map (against which it was not refined; FSC_{half2}; red line) is also shown. **d**, Flow chart for cryo-EM data processing and structure determination of the ABCG2_{EQ}–E₁S complex. **e**, Full view of the final CryoSPARC B-factor-sharpened map of ABCG2_{EQ}–E₁S, coloured by local resolution in Å, as calculated by ResMap with the clipping plane in the middle of the molecule. **f**, Angular distribution plot for the final reconstruction.



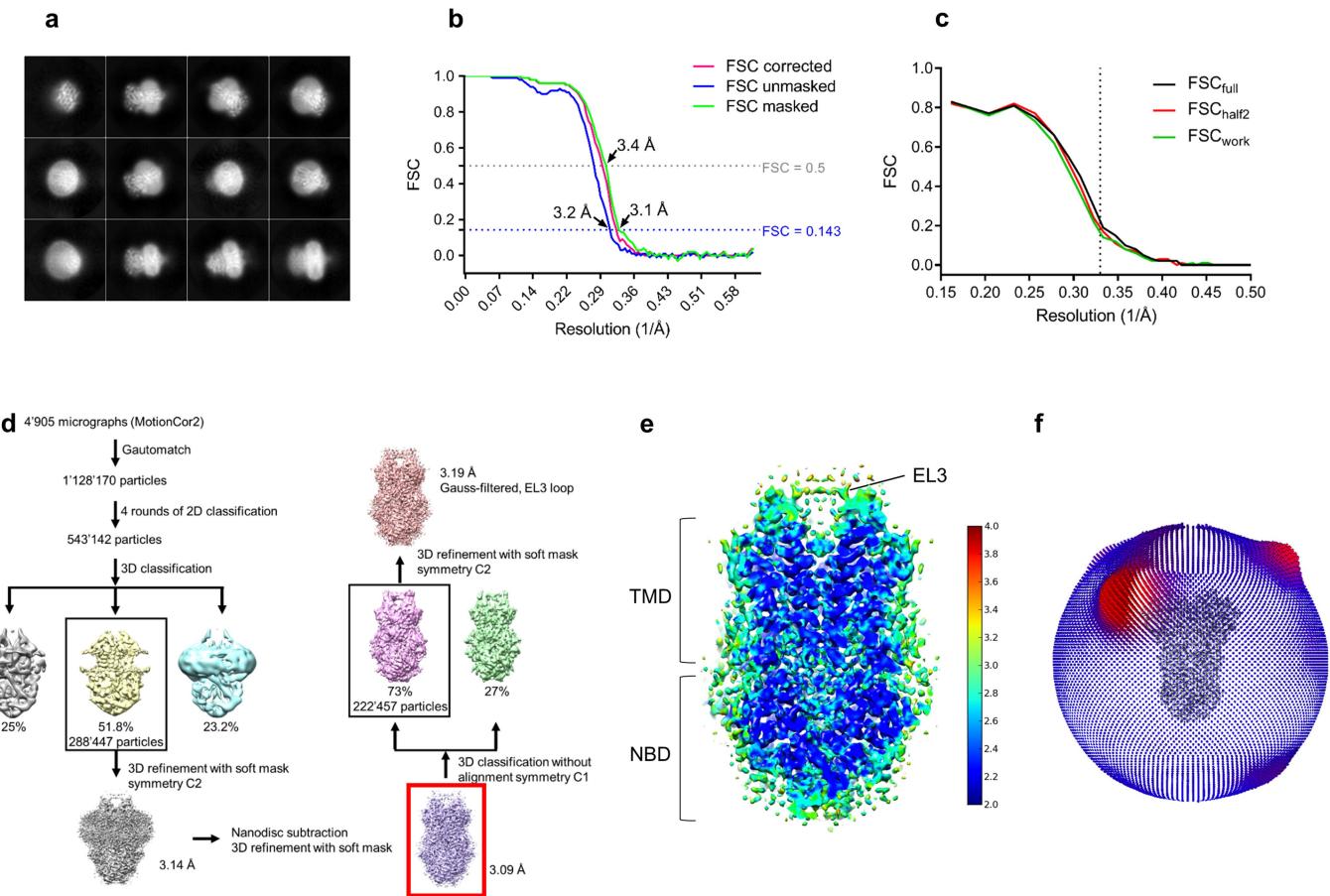
Extended Data Fig. 4 | Fit of the models to the densities. **a**, Fit of the TM helices of the final model of the ABCG2_{EQ}–E₁S TMD to the post-processed and masked C2 map from CryoSPARC. A region of up to 2 Å around the atoms is shown. **b**, The fit of one E₁S molecule (pink or turquoise sticks) in two possible orientations, flipped by 180°, docked into the C2-symmetrized substrate density of the final model of ABCG2_{EQ}–E₁S. The contour level has been reduced by comparison with Fig. 2a to show

the strongest density at the core of the polycyclic rings. **c**, As for **b**, but showing the fit of one E₁S into the electron microscopy density of the post-processed and masked C1 map from CryoSPARC. **d**, Fit of the TM helices of the final model of the ABCG2_{EQ}–ATP TMD to the post-processed and masked C2 map from RELION. A region of up to 2 Å around the atoms is shown.



Extended Data Fig. 5 | Purification and functional analysis of mutants.
a, Analytical SEC profiles of the detergent-purified wild-type and substrate-binding cavity mutants used to make proteoliposomes for functional assays. **b**, ATPase rates of the liposome-reconstituted wild-type and mutant proteins in the presence of 0–300 μM E₁S. Each point represents the mean rate derived from technical replicates (same batch

of liposomes) and error bars show the standard deviation. For G2, $n=6$, except in the case of 0 and 200 μM E₁S, for which $n=9$. For the mutants, $n=3$. **c**, Table showing the EC₅₀ of E₁S ATPase stimulation determined after normalizing the curves in **b** with the error of the fit (standard deviation) shown.



Extended Data Fig. 6 | Cryo-EM map generation, data processing and atomic-model refinement of ABCG2_{EQ}-ATP. **a**, Twelve representative 2D class averages of the final round of 2D classification, sorted in decreasing order by the number of particles assigned to each class. **b**, FSC from the RELION auto-refine procedure of the unmasked half-maps (blue), the half-maps after masking (green), and the half-maps after masking and correction for the influence of the mask (pink). A horizontal dotted line (blue) is drawn for the FSC = 0.143 criterion. For both the unmasked and the corrected FSC curves, their intersection with the FSC = 0.143 and the FSC = 0.5 lines are marked by arrows, and the resolutions at these points are indicated. **c**, FSC curve of the final 3.09 Å refined model versus the map against which it was refined (FSC_{full}; black line). FSC

curves of the final refined model with introduced shifts (mean value of 0.3 Å) versus the first of two independent half-maps (half-map 1, against which it was refined; FSC_{work}; green line) or the same model versus the second independent half-map (against which it was not refined; FSC_{half2}; red line) are also shown. **d**, Flow chart for cryo-EM data processing and structure determination of the ABCG2_{EQ}-ATP complex. The map used for model building is indicated by a red square. **e**, Full view of the RELION local-resolution-filtered map of ABCG2_{EQ}-ATP, coloured by local resolution in Å as calculated by ResMap, with the clipping plane in the middle of the molecule. **f**, Angular distribution plot for the final reconstruction.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	ABCG2 _{EQ} -E ₁ S	ABCG2 _{EQ} -ATP
	(EMD-0196, PDB 6HCO)	(EMD-0190, PDB 6HBU)
Data collection and processing		
Magnification (nominal)	61,610 \times (165k \times)	61,610 \times (165k \times)
Voltage (kV)	300	300
Electron exposure (e $^-$ /Å 2)	2.0	2.0
Defocus range (μm)	-0.7 to -2.8	-0.5 to -3.3
Pixel size (Å)	0.812	0.812
Symmetry imposed	C2	C2
Initial particle images (no.)	168,184	1,128,170
Final particle images (no.)	42,790	288,447
Map resolution (Å)	3.58	3.09
FSC threshold	0.143	0.143
Map resolution range (Å)	308.6-3.58	308.6-3.09
Refinement		
Initial model used	PDB 6ETI	PDB 6ETI
Model resolution (Å)	3.58	3.09
FSC threshold	0.143	0.143
Model resolution range (Å)	207.0-3.6	207.0-3.1
Map sharpening B factor (Å 2)	-82.6	-136.0
Model composition		
Nonhydrogen atoms	12,338	9,144
Protein residues	1576	1168
Ligands	24	64
B factors (Å 2)		
Protein	140.93	22.25
Ligand	118.32 (E ₁ S)	9.3 (ATP) 4.1 (Mg $^{2+}$)
R.m.s. deviations		
Bond lengths (Å)	0.009	0.009
Bond angles (°)	1.010	1.087
Validation		
MolProbity score	1.82	1.61
Clashscore	7.26	4.34
Poor rotamers (%)	0.45	0.00
Ramachandran plot		
Favored (%)	92.20	94.62
Allowed (%)	7.80	5.03
Disallowed (%)	0.00	0.35

For the ABCG2_{EQ}-E₁S structure, only the variable domain of 5D3-Fab was modelled.

CAREERS

INTERVENTION Why PIs must recognize and support trainees with depression **p.433**

WELL-BEING For advice and support on mental health go.nature.com/naturejobs

 **GOT A STORY?** Contact the editors naturecareereditor@nature.com

CHRIS ENSOUL



Clinical psychologist Kate Baecher has combined her love of mountaineering with scientific expertise to study risk, fear and survival among mountain athletes.

INTERDISCIPLINARY RESEARCH

Turning a passion into a job

Outside pursuits needn't be just side gigs. Scientists are fusing disciplines to form new fields.

BY EMILY SOHN

Indre Viskontas took piano lessons as a child and made her opera debut at age 11. But her mother, a professional conductor, told her that music did not pay well. So Viskontas, who often listened to the opera singer Maria Callas while doing homework, decided to pursue science instead, earning an undergraduate degree in psychology and French literature at the University of Toronto, Canada, and a PhD in cognitive neuroscience at the University of California, Los Angeles. During a year in London, she took singing lessons that she continued during her PhD, when she also sang opera.

Viskontas saw neuroscience as a stable career choice that might offer ideas about how to better embody roles in operatic performances. But after years of alternating her focus between

science and music, she found a way to combine the two, by applying neuroscience to musical training. She now works as an opera singer and cognitive neuroscientist, with positions at the University of San Francisco, California, and the San Francisco Conservatory of Music.

Scientists who have successfully crafted a research career out of their non-academic passions and talents say that persistence and patience are key, especially when trying to merge two professional paths that might not seem obviously connected. Melding worlds can be unsettling, and it takes time and creativity to persuade funders and advisers that the work is worthwhile.

But those who have done so say that focusing a research lens on their life's passions has expanded both their personal horizons and scientific goals in an academic landscape that

is becoming increasingly interdisciplinary. Viskontas' research projects include teaching people with cochlear implants how to sing. "My life is like this DNA double helix constantly turning itself over," says Viskontas, adding that she has noticed a drive towards innovative solutions in science and a demand for a more scientific approach in the arts. "I feel like I'm a bridge between these worlds. And we are entering an era where that overlap is more celebrated."

FORGING A PATH

It often takes perseverance to find ways to study one's personal interests, especially if those interests don't already belong to established research departments, says Vanesa España-Romero, an exercise physiologist at the University of Cádiz in Spain. She got hooked on rock climbing while she was at school ▶

► and chose to pursue sports science, partly as a way to learn how to be a better climber.

But when she began studying elements of climbing fitness for her PhD, such as handgrip strength and percentage of body fat, nobody else in Spain was studying the sport. She had to create tools and formulas while explaining what she was doing to colleagues and supervisors. Funding was impossible to get, she adds, because grant providers didn't understand climbing or see any reason to study it. "People felt I was doing something weird," she says.

Instead of giving up, Espa  a-Romero recommends, go for long-term thinking and creative strategies for forging new research paths. She sought funding for research into the promotion of physical activity and health. Then, she applied existing research tools to her climbing studies. It's a strategy she still uses, although increasing interest in the sport has made it possible for her to get a little money for climbing-specific studies. "I think the key thing for me was the persistence," she says. "If you love what you do, go for it. But you need time."

During the time it can take to work out how to combine science with an outside interest, it might be necessary to pursue both in tandem. Good organizational skills can help researchers to juggle two identities at once, says neuroscientist Peter Vuust, who is director of the Center for Music in the Brain at Aarhus University in Denmark. He also teaches music at the Royal Academy of Music in Aarhus, and is a bassist. His research addresses questions about how the brain processes music, with projects such as the use of music in health care.

Vuust started playing music professionally when he was 16, but studied French and music as an undergraduate, mathematics for his master's degree and neuroscience for his PhD. Even now, as a working scientist, Vuust plays music every morning at 6:30 for up to an hour and a half. It's meditative time for him that helps him to maintain a performance schedule of 60 concerts a year.

To keep side interests alive while working towards a science degree, Viskontas recommends being strategic about institutes and supervisors. She sought advisers who were supportive of students who follow outside interests and maintain strong publication records. She stuck with a supervisor who allowed her to work whatever hours she wanted. Another mentor attended her musical performances, and asked about her music before asking about her data.

FINDING FUNDING

A strategic approach also applies to funding, Viskontas says. To pursue her music while studying neuroscience without objections from her graduate-programme leaders, she spent a lot of time researching and applying for independent grants. Her search included family foundations and opportunities that were aimed at highly specific groups. She received one award for scientists who were pursuing extracurricular activities. The approach allowed her to self-fund her PhD, giving her financial independence from her institution, and the ability to take a break from her PhD for up to one month each summer. During that time off, she performed in operas in locations such as Italy and Canada. "I would go through literally hundreds of scholarships and grant opportunities," she says. "Then I would send a letter to them and say, 'Hey, I'm studying this really fascinating thing.' You can target your letter to fit the foundation's mission."

Vuust took a different approach to the same need for freedom. For two years, he worked every day on applying for a major grant from the Danish National Research Foundation, which is given to about ten scientists once every three years. He didn't get it, and had to rely instead on smaller grants. In 2014, with a polished application, he got the grant, allowing him to focus on his research and his music without worrying too much about the need to constantly seek more money.

Enduring a long wait before finding a way

to combine science with other interests isn't necessarily a bad thing, adds Kate Baecher, an independent clinical and performance psychologist in Sydney, Australia. For years, she focused her research on trauma among veterans, while pursuing a love for mountaineering in her own time. Only in the past year or two has she begun to study risk, fear and survival among mountain athletes on expeditions. Combining her scientific expertise with her love of mountains made sense only after she had established herself as a psychologist and developed the skills to ask and address questions she wanted to answer. "I wouldn't have been ready to do it at the begin-

"I saw a way in which my neuroscience training could benefit musicians."

ning of my career, when I was still learning," she says. "Now, I have the professional and personal maturity to tackle it."

To avoid burnout, Baecher recommends drawing clear boundaries between work and play hours. She still climbs mountains with friends, purely for the joy of it, including a month-long trip earlier this year to Pakistan. And she pursues research projects that have nothing to do with mountaineering. "Professional balance is really important," she says.

MERGING SCIENCE AND THE ARTS

Insecurities can grow while trying to combine two disparate identities, but pay-offs often emerge down the line. Now that Vuust has embraced both music and neuroscience, he feels more confident in his music, because there is less at stake. "When I got a career as a brain scientist, that gave me two legs to stand on," he says. "If I didn't play as well in a certain gig, that didn't bother me as much as it used to."

Some scientists find that studying their passions can enhance performance. Using her work, Espa  a-Romero is now able to apply evidence-based performance strategies to climbing, and she helps other climbers to do the same. When her muscles are under strain during a climb, for example, she is able to assess how many seconds of rest are needed before making another move. In 2014, Vuust received his first nomination for a Danish Music Award for an album that incorporated insights from his brain research to achieve maximum emotional impact in listeners. At a key moment in the title song, he used a melody note and an unexpected minor chord to coincide with a crucial word, hoping to tap into the brain's system of musical prediction. "To me, it really sounds like opening the window and feeling the wind blowing from a cold November day," he says. "The idea is to musically emphasize the emotions related to coming of winter."

Pursuing both science and outside interests can also lead to new discoveries and ideas, says Viskontas. After finishing her neuroscience PhD, she chose to do a graduate degree in music. As she studied and practised, she began to recognize misconceptions in conventional



Danish neuroscientist Peter Vuust heads a lab, teaches music and plays his bass in 60 concerts a year.

MADS BJOERN CHRISTIANSEN

practice methods, including a reliance on long hours of rote learning. Instead, her understanding of memory in the brain suggested that shorter but more varied sessions would challenge the brain to learn faster. The realization opened up a new career path, combining science with music. "All of a sudden, I saw a way in which my neuroscience training could benefit musicians and still be interesting to me," she says, adding that she now accomplishes in a 30-minute practice session what used to require 4–8 hours of work. "I could hack my practice time with neuroscience."

Turning a scientific lens onto outside interests sometimes helps to create new fields of science. Emma Redding began her career as a contemporary dancer and later started teaching, which led to an interest in how training methods could help dancers to meet the high physical demands of dancing. But when she did a master's in sports science, she had to learn about the biomechanics and physiology of sports such as rugby and football. There was no one to teach her about dance. In 2000, she wrote the first master's degree on dance science.

Now head of dance science at Trinity Laban Conservatoire of Music and Dance in London, where 25 graduate students enrol each year, Redding has watched the field grow to include as many as 10 undergraduate and graduate dance-science programmes around the United States and Europe. But her choice to merge science and dance required a leap of faith, and she still faces scepticism from people who think that dance is an art form that doesn't belong in the realm of science. "I suppose I was attempting to study something that didn't exist," she says. "That's why I had to start with it as an interest or hobby. Then when I got qualifications in science, I was able to start trying to develop the field."

Studying one's passion can lead to new opportunities, Viskontas adds. She has been using her performance skills to communicate science through online lectures and as host of two podcasts and a television series. In addition to neuroscience research, she works on a couple of musical projects a year, including an upcoming performance of a psychological thriller with a feminist twist that is being written for her voice. This year, she directed a version of an opera called *The Man who Mistook his Wife for a Hat*, based on an essay by the late neurologist Oliver Sacks, who was once her mentor.

Researching any type of science requires intense dedication and energy, Vuust says, adding that the best scientists are those who study what they love. "In order to be a really good researcher, it has to be a passion," he says. "What you do has to be fun." ■

COLUMN

Lab listener

James Turner extols the value of mental-health first aid.

The Francis Crick Institute in London now has around 40 accredited mental-health first aiders. The two-day training course is run by our occupational-health nurse, covering conditions such as anxiety, depression, eating disorders and psychosis.

I volunteered because there weren't yet any scientific group leaders among the first aiders. We should have them at all levels of the organization, and managers should be exemplars. Like anyone else, we experience mental-health problems. I had them in the past, and my experience taught me that things can deteriorate quickly — and that early intervention is key.

I studied psychiatry during my medical degree. We focused on diagnosis and therapy, but there was less emphasis on listening skills.

COURSE BASICS

The Crick's training course teaches you to listen in a non-judgemental way, to pay attention to negative signs and not be afraid of asking difficult questions, such as, "How are you? I've noticed you're not quite yourself." One point made during the training that isn't always captured in textbooks is that two people can experience the same mental illness very differently. Another is that recovery is possible, but you have to give it time.

There's also a strong emphasis on using the right terminology and avoiding inappropriate language. The phrase "committing suicide", for example, implies that someone has performed a crime. "Completing suicide" or "taking one's own life" are more appropriate.

The Crick's mental-health first-aid network started in 2016 when the institute opened. At first, most volunteers were women. It took a while to get men on board, but now the network has equal numbers. Diverse representation is important because some mental-health conditions affect men and women differently. For example, in the United Kingdom, three-quarters of people that complete suicide are men, according to the Samaritans' 2018 Suicide Statistics Report (see go.nature.com/2rpp8du). I work on sex differences as part of my research, and so find that statistic interesting. I hope that by making more men aware of and engaged in these initiatives, we might understand why they're less likely than women to self-report mental-health issues.

My advice to someone with a mental-health problem is to remember that you are one of many going through this. The World Health Organization notes that one in four people



globally will be affected at some point in their lives. But don't accept it as the norm.

The help we offer is confidential. Our contact details are available on our intranet and on notice boards throughout the building. We are 'signposters', there to listen, not to judge, and to refer people to an appropriate service. This could be their own general practitioner, or Health Assured — the Crick's external-assistance-programme provider. External charity organizations, such as Samaritans, Mind and SANE, offer more sources of support.

The first aiders use a WhatsApp group to communicate and support each other. There is also a group debriefing session every eight weeks, and the Crick offers half-day courses to help individual employees look after their own mental health and to manage stress more effectively.

WORKPLACE CULTURE

Academic science is a fantastic but challenging career choice. Competition for jobs is huge, expectations from scientific journals are high and a scientist's role is ever-changing. Alongside research, we teach, raise funds and engage with the public and the media. Juggling these responsibilities can be tough.

Some scientists say that stress is part of the job, and wear it like a badge of honour. I want to debunk that myth. Mentoring schemes and health-awareness events, which we have at the Crick can provide scientists with day-to-day support. We senior scientists should also coach trainees on how to cope with the pressures of a research environment. I strongly believe that with great mental health comes great science. We should all get on board with this message. ■

James Turner is a senior group leader at the Francis Crick Institute in London, where he runs the Sex Chromosome Biology Lab.

SAY IT WITH MASTODONS

A project from the heart.

BY MARISSA LINGEN

It is completely implausible that people should fall in love with each other, but of course they do. I mean, we do. Even I do, apparently, although it makes no sense to me; the perfectly logical reasons why I should like you not being enough, although I can see no reason that they shouldn't be, that I should be weirdly tender towards you in addition to all of the ordinary human respect and, damn it, *liking* that you inspire in me. That I should not only think of you often but smile a weird wobbly smile when I do. That I should care so much, all of a sudden, about little things that make you happy, about how your week is going, about the way you read things over the top of your glasses and the way your eyes crinkle up when you smile. It's far more specific than a general fondness, very intense. It doesn't make any sense even though most people do it. It's pretty weird, and I really don't know what to do about it.

So I made you some mastodons.

I don't think this is just a me thing. I think it's an us thing. As much as I understand falling in love as a thing, I think part of it is that there are so many us things. Like mastodons. Or, more to the point, like soil restoration on the northern Great Plains.

We've talked about it so many times, sometimes in our labs and sometimes in charming cafés with thoughtfully concocted beverages, and sometimes in tree-lined parks. We talk a lot. We talk about this. About how we both grew up in small towns in the Dakotas — you outside Watertown, me in Wahpeton, close enough to marvel at now that we're not there. About how hope rose in us, in strange half-understood little kid ways, when the grazing programme to restore the soil really took off, when the crop yield went up and the air smelt of wet, green, growing things. Even though we didn't know each other then, we remembered how it was. We knew.

So we both knew the difference, when the blackleg hit. When climate change made its season longer and mutation made it rage

on. We knew what those farms could have been, what they were struggling to be in the newfound heat and tornadoes. They struggled, they stumbled, and we came of age — separately, but together — with the dusty chemical smell of failure around us.

with their ancestors, that's small and manageable. But when you don't expect mastodons, it's still a lot. Kind of like falling in love when a person didn't plan to, I guess.

It's fun to collaborate, so ... I left it to you to figure out what trees the farmers should

plant at their field borders, for the mastodons to browse when they're not grazing. Almost anything would be okay, but surely there are some that they'll like best. Mastodons like soft shoots and fresh leaves in the spring. They're surprisingly focused on tenderness, although I shouldn't be surprised by anybody that way after recent developments.

They're still good grassland maintainers, though, I'm pretty sure. They should be. The lab tests show their manure to be rich and fertile and nitrogen-fixing. They've had a lot to process, these mastodons. I can relate.

I made them blackleg-resistant, at least as much as anyone can ever be sure they're resistant to anything. They haven't responded to any of the fatal strains that are common now, so we can hope for at least a reprieve. Some quiet time to do their job without worrying

about new developments. Wouldn't that be nice.

Anyway, here's the key. I hope you want to visit them, as I made them for the place we came from but really mostly for you. You can get right into the pens with them if you're careful. I know their long tusks look scary, but they're intensely affectionate when you get in past their guard. Not that I know what that's like.

I know this is going to feel sudden, even though it took almost two years just to gestate them, so I guess I've been squirming about this for some time. I'm sorry if I'm making you squirm alongside me — that's one of the things I actually *don't* want to share. It's okay if you don't want to talk about the love part. It makes me uncomfortable too. We can just talk about the mastodons. I like them best.

Well. Best except for you. ■

Marissa Lingen has published more than 100 short stories in venues such as *Analog*, *Lightspeed* and *Tor.com*.

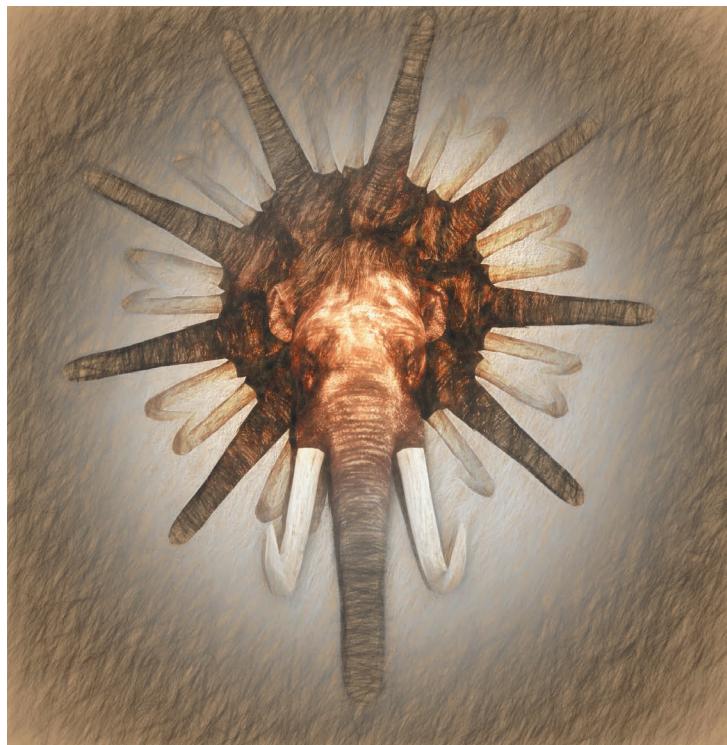


ILLUSTRATION BY JACEY

We both went east for school, like so many people. We both turned away. Neither of us could help looking back, and then we found each other. And then all the conversations, while the blackleg spread to cow and bison. While no one could figure out what to do about it.

And then I fell in love with you and I didn't know what to do, and I thought of where we're from. I thought of the things we'd turned over in conversation. The tread of cattle on the plains, the natural fertilizer, the way it had all gone so right before it went so wrong. I know we're both focused on conservation, but the conservation attempts weren't getting the prairie anywhere. We needed another, more radical, solution and also being in love with you turned my brains upside down and then right side up again.

And so I thought: mastodons.

Just little ones. Not much over half a tonne. Compared